World Journal of Innovation and Modern Technology, Vol. 8, Issue 7 (Jul) 2025
ISSN 2682-5910
DOI: 10.53469/wjimt.2025.08(07).09

D-SELACFP: Visible-Infrared Person Re-Identification via Local Attention and Cross-Modal Feature Perception Enhancement

Li Fan

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

Abstract: ***Visible-Infrared Person Re-Identification (VI-ReID) faces significant challenges such as difficult feature matching and large pose variations, leading to low cross-modality image matching accuracy. Existing approaches typically extract features directly from raw images and embed features from different modalities into a common space to learn shared representations. However, they often neglect the interference of noise and fail to fully utilize identity-discriminative information within modality-specific features, resulting in weak cross-modality invariant feature extraction and interference from irrelevant noise during feature matching. To address the issues of modal noise interference and feature matching difficulty in VI-ReID, this paper proposes a dual-stream neural network framework designed to reduce modal noise and enhance perceptual features. The framework incorporates a Local Attention Module (LAM) and an Inter-modal Feature Perception Enhancement Module (I-MFPE). These modules work synergistically to mine salient features both within and across modalities, while effectively leveraging the identity-discriminative information inherent in modality-specific features. Specifically, the LAM focuses on discriminative part-level features and suppresses background noise interference, thereby effectively extracting both modality-shared and modality-specific features while reducing the impact of irrelevant information. The I-MFPE module enhances the ability to extract shareable features from heterogeneous images by optimizing fine-grained feature representations in both channel and spatial dimensions, simultaneously mitigating the influence of modal differences on matching. The proposed method effectively alleviates noise introduced by factors such as viewpoint variations and background clutter, enhancing the discriminative power of cross-modality pedestrian features and providing a more robust feature representation for VI-ReID. Experimental results demonstrate the significant advantages of the proposed method in suppressing noise interference and improving feature matching performance.***

Keywords: Visible-Infrared Person Re-Identification, Local Attention Module, Modal Noise, Feature Perception Enhancement.

1. INTRODUCTION

Person Re-identification (Person Re-ID) refers to the technology of matching and retrieving target individuals across non-overlapping camera views [1]. It holds significant application value in public security domains such as intelligent surveillance and behavior tracking. Surveillance scenarios often suffer from low camera resolution and restricted viewing angles, making it difficult to capture high-quality facial images. Consequently, Re-ID has become a crucial alternative technology when facial recognition fails. However, traditional visible-light-based Re-ID methods face challenges under low-light conditions such as nighttime, as visible-light cameras struggle to capture discernible pedestrian appearance features in such environments [2].

To address this limitation, modern surveillance systems typically employ Visible-Infrared dual-mode switching mechanisms, automatically activating infrared imaging at night to obtain clearer images. Nevertheless, significant modality differences exist between infrared and visible-light images, causing performance degradation in traditional single-modality Re-ID methods. To overcome this challenge, researchers have proposed Visible-Infrared Person Re-Identification (VI-ReID) technology, aiming to bridge the cross-modal gap and enhance the accuracy of pedestrian identity matching and retrieval across different imaging modalities.

In the field of person re-identification, mainstream deep learning-based approaches primarily include representation learning, metric learning, and Generative Adversarial Networks (GANs). VI-ReID exhibits significant differences from traditional Re-ID and facial recognition technologies, as its core research objective is to mitigate the feature disparity between visible and infrared modalities. Early research, predominantly based on the RegDB dataset [3], focused on coarse-grained feature extraction while overlooking critical fine-grained information, resulting in limited matching accuracy. With advancing research, Hao et al. [4] proposed a feature segmentation-based method for extracting local details, effectively alleviating spatial misalignment issues in

pedestrian images. Cheng et al. [5] designed a dual-stream fusion network that enhanced model performance through a multi-granularity partitioning strategy, though its complex architecture substantially increased computational costs. Addressing this issue, Liu et al. [6] innovatively employed a skip-connection structure in convolutional neural networks to achieve effective fusion of coarse- and fine-grained features, significantly improving the model's perception of detail. To further optimize feature representation, Wang et al. [7] proposed a multi-branch matching network, though limitations remained in reducing the inter-modal distance. Consequently, Zhu et al. [8] designed a Heterogeneous Center Loss function, effectively enhancing the discriminative power of pedestrian features. Research in 2021 achieved important breakthroughs: Fu et al. [9] systematically analyzed network architecture and demonstrated that the appropriate separation of Batch Normalization (BN) layers is critical for improving cross-modal matching accuracy. Concurrently, Ye et al. [10] proposed an innovative Random Channel Enhancement method, which significantly boosted cross-modal matching performance and could be seamlessly integrated into existing models without structural modifications. These studies have provided new insights and methodologies for advancing VI-ReID technology.

Recent years have witnessed the remarkable advantages of attention mechanisms in deep learning, particularly in information processing and feature selection, demonstrating human visual perception-like capabilities: Hu et al. [11] pioneered Squeeze-and-Excitation Networks (SENet), which adaptively recalibrates channel feature responses by modeling interdependencies among feature channels. Woo et al. [12] further proposed the Convolutional Block Attention Module (CBAM), innovatively combining channel attention with spatial attention using a dual-path aggregation strategy of global average pooling and max pooling. In the Re-ID domain, attention mechanisms have been widely applied for feature refinement [13]. Addressing the limitations of traditional CNN models in capturing long-range dependencies, Liu et al. [14] designed a dual-branch self-attention network that effectively integrates global context with local detail features. Li et al. [15] proposed an attribute reasoning-based attention framework that precisely locates attribute features through a spatial-channel cooperative attention mechanism. Zhao et al. [16] extended spatial-channel attention to cross-modal scenarios, significantly enhancing feature discriminability. However, these methods face computational efficiency bottlenecks when processing high-dimensional cross-modal data: the global interactions of self-attention incur high computational complexity, while the cascaded structure of spatial-channel attention also substantially increases computational overhead.

In summary, Visible-Infrared cross-modal person re-identification research faces two key challenges: (1) significant modality differences leading to difficult feature alignment, and (2) insufficient exploration and ineffective fusion of multi-granularity features. To address these issues, this paper proposes a cross-modal person re-identification network based on Local Attention and Inter-modal Feature Perception Enhancement. The network adopts a joint yet independent dual-path attention learning strategy for more effective salient feature extraction and achieves efficient cross-modal feature fusion through a non-local interaction mechanism. The framework employs a coarse-to-fine hierarchical learning strategy to systematically mine discriminative identity features across different scales, significantly enhancing feature representation distinctiveness. The proposed cross-scale feature perception strategy effectively mitigates noise interference from complex backgrounds surrounding pedestrians through a multi-scale feature interaction mechanism. Experiments demonstrate its significant effectiveness in enhancing cross-modal feature robustness. The introduced local attention mechanism incorporates frequency-domain feature compression technology, adaptively enhancing key frequency components crucial for identity recognition while suppressing non-discriminative interference factors like background noise through frequency-domain filtering. This approach optimizes computational efficiency while preserving feature representation richness. By systematically addressing critical issues in cross-modal feature fusion and multi-granularity feature mining, this study provides novel technical insights for Visible-Infrared person re-identification.

2. METHOD

2.1 D-SELACFP Framework

To address the challenges of significant modality differences, abundant background interference, and insufficient feature representation in cross-modal person re-identification, this paper proposes a deep network framework based on a Local Attention Module (LAM) and an Inter-modal Feature Perception Enhancement Module (IFPEM), as illustrated in Figure. 1. This network employs a dual-stream ResNet-50 architecture where the shallow convolutional modules (Layer 0) for the visible-light path and infrared path utilize independent parameters to accommodate modality-specific characteristics. In contrast, the deep convolutional modules (Layers 1~4) share parameters to learn modality-invariant features. The framework comprises three core components: backbone

feature extraction network, Local Attention Module (LAM), and Inter-modal Feature Perception Enhancement Module (IFPEM), responsible for fundamental feature extraction, key region focusing, and cross-modal feature alignment respectively.

Inspired by multi-granularity feature learning [18], this study similarly designs a multi-granularity non-local fusion framework. This framework first employs the LAM module to suppress noise and enhance discriminative features, followed by feature enhancement in spatial and channel dimensions. An end-to-end feature learning strategy is adopted, integrating shallow and deep network features through cascade fusion to extract multi-granularity feature representations. Specifically, features from both modalities are initially processed through 1×1 convolution, Batch Normalization (BN), ReLU activation, and max pooling operations to preserve critical features while enhancing network expression capability. Subsequently, through the Inter-modal Multi-scale Feature Perception Enhancement strategy (I-MFPE), multi-scale feature fusion attention mechanisms and feature extraction layers process the features. Finally, channel attention features and spatial attention features undergo deep fusion to generate robust cross-modal fusion features.

Regarding feature optimization: Generalized Mean Pooling (GeM) dynamically adjusts pooling behavior to better capture feature map characteristics; Cross-entropy loss and orthogonal loss enhance feature discriminability and consistency; Heterogeneous center triplet loss effectively extracts and fuses discriminative information from modality-specific features. The synergistic effect of these technologies achieves deep feature optimization and significant performance improvement.

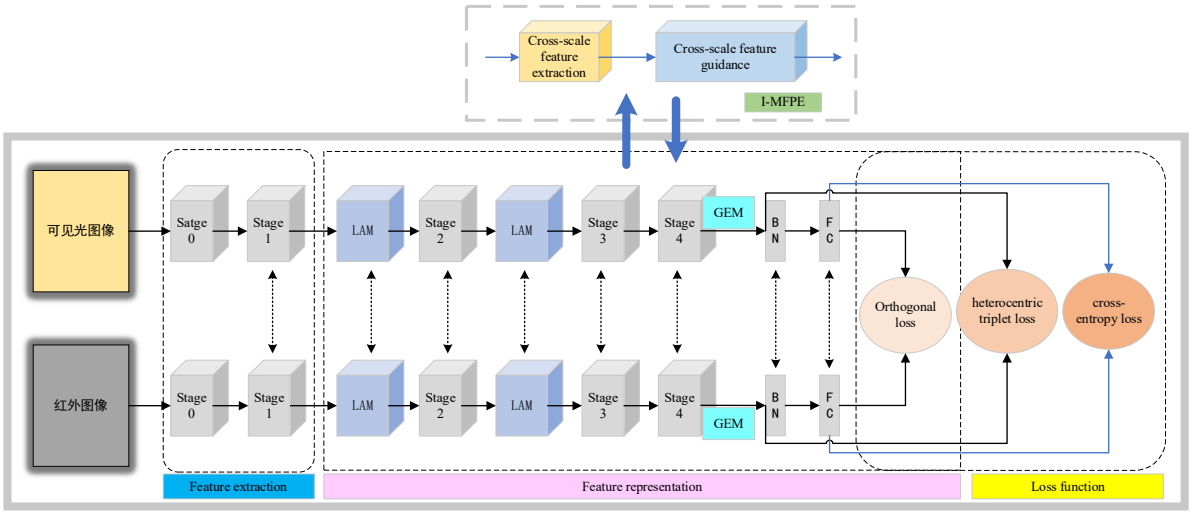

Figure 1: D-SELACFP dual-stream network framework

2.2 LAM Module

To enhance the model's ability to extract discriminative pedestrian features while suppressing cross-modal interference from background noise, this paper proposes a Local Attention Mechanism (Local Attention Module, LAM). As shown in Figure. 2, traditional channel attention mechanisms typically rely on Global Average Pooling (GAP) to obtain channel-level global information. However, when processing cross-modal data (e.g., Visible-Infrared images), GAP tends to lose critical frequency information, resulting in limited expressive capability of modal features.

Addressing this limitation, this paper employs Two-Dimensional Discrete Cosine Transform (2D-DCT) to optimize the feature encoding process. Unlike 1D-DCT, 2D-DCT is specifically designed for image data. It decomposes spatial domain features into cosine basis functions of different frequencies, enabling efficient modeling of local textures and spatial correlations. This transformation naturally possesses advantages in image compression and denoising while preserving significant frequency components, making it more suitable for cross-modal feature enhancement.

Specifically, the LAM module first divides the input feature map into n local regions. It then independently performs 2D-DCT transformation on each partition to generate frequency-domain feature representations. This process is implemented through three steps: feature map partitioning, frequency-domain transformation, and

feature reconstruction. Through frequency-domain local modeling, this method significantly enhances the ability to capture discriminative pedestrian features such as clothing textures and contours in cross-modal scenarios, while effectively mitigating interference caused by lighting variations and background clutter.

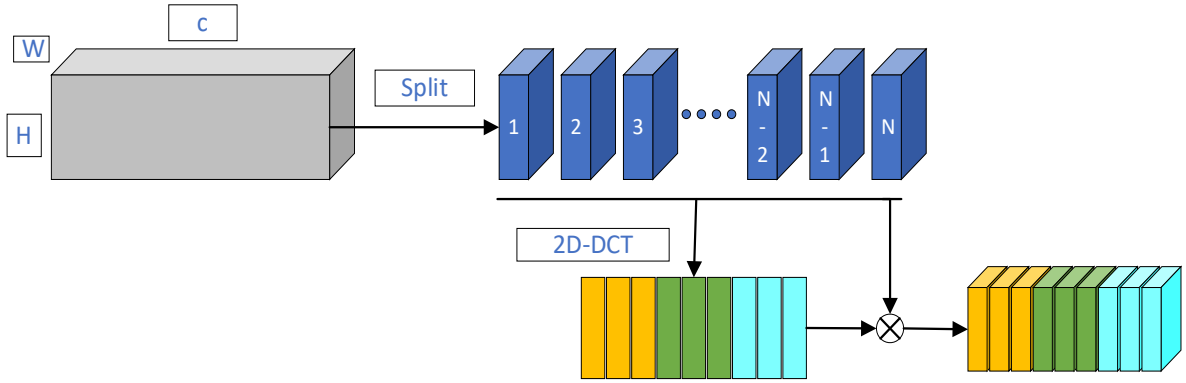

Figure 2: LAM module structure diagram

2.3 I-MFPE Module

Deep To effectively suppress background interference around pedestrians and enhance the salient representation of pedestrian regions, this paper proposes an Inter-modal Feature Perception Enhancement module. As illustrated in Figure. 3, this module deeply explores the discriminative feature differences between pedestrian targets and complex backgrounds through a multi-scale feature fusion mechanism, thereby enhancing the feature separability between pedestrian regions and background regions.

Specifically, given the input low-level features (containing local detail information) and high-level features (containing semantic boundary information), the module first adjusts the channel dimensions through two independent 1×1 convolutional layers, unifying the feature channels to 64 and 256 respectively. Subsequently, bilinear upsampling is performed on the high-level features to match the spatial dimensions of the low-level features. After feature concatenation, sequence transformation consisting of two 3×3 convolutional layers and one 1×1 convolutional layer is employed for feature fusion, ultimately outputting cross-scale semantic features with a channel count of 1.

The proposed cross-scale feature perception strategy comprises two core components: the cross-scale feature extraction module and the cross-scale guided feature module. Experimental results demonstrate that this strategy can effectively enhance the network's reasoning ability for multi-modal data. Particularly under scenarios with significant background interference, it maintains stable feature extraction performance.

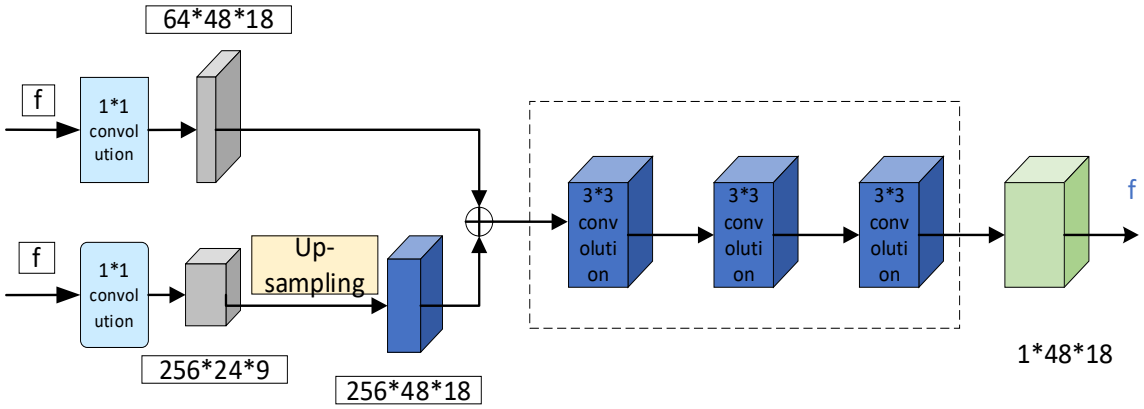

Figure 3: I-MFPE module structure diagram

3. EXPERIMENTS AND RESULTS ANALYSIS

3.1 Datasets and Evaluation Metrics

This experiment evaluates model performance on two public Visible-Infrared person re-identification datasets

(SYSU-MM01 and RegDB): SYSU-MM01[20] is a large-scale cross-modality dataset containing 491 pedestrian identities captured by 4 visible-light cameras and 2 infrared cameras, comprising 287,628 visible-light images and 15,792 infrared images, with its training set containing 34,167 images of 395 identities and the test set containing the remaining 96 identities, supporting both All-search and Indoor-search testing modes; the RegDB dataset [3] collected by Visible-Infrared dual-camera systems contains 412 pedestrian identities (each providing 10 visible-light and 10 infrared images) randomly split into training and test sets at a 1:1 ratio, supporting bidirectional VIS→IR and IR→VIS retrieval modes. Evaluation metrics include Rank-k (R-k) indicating correct match probability in top-k retrievals and mean Average Precision (mAP) measuring positive sample distribution (higher values indicate top-ranked matches), collectively assessing retrieval accuracy and robustness in cross-modal person re-identification.

3.2 Experimental Settings

The experiment is implemented under the PyTorch framework using ImageNet-pretrained ResNet-50 as the baseline model (Baseline), where a dual-branch structure extracts modality-specific and modality-shared features constrained by a modal purification restorer and modal consistency guider; models are trained with 384×128 pixel inputs using random horizontal flipping, random cropping, and random erasing for data augmentation, where each training batch contains 12 identities (10 images per identity, totaling 120 images) with learning rate decaying to 0.1× original at epochs 60 and 100; for the RegDB dataset's limited samples, the network is modified by training only up to Stage 3, inserting LAM modules before Layer2/Layer3, and implementing Inter-modal Feature Perception Enhancement (IFPEM) between Layer2-Layer3—the 150-epoch training process converges through joint optimization of cross-entropy loss and triplet loss, with final evaluation using modality-shared features.

3.3 Comparison with State-of-the-Art Methods

To verify the effectiveness of the method proposed in this paper, we conducted comparative experiments with the existing mainstream methods on two public datasets, SYSU-MM01 and RegDB. The experimental objects include three representative methods: the method based on feature alignment (MPANet [25] et al.), the method for processing image noise (NFS [22], DART [24] et al.), and the method for enhancing feature information (Hi-CMD [21] et al.). The experimental results show that the method proposed in this paper has achieved significant improvements in multiple evaluation indicators. In the full search mode of the SYSU-MM01 dataset, Rank-1 and mAP reached 82.77% and 79.91% respectively; In the indoor search mode, Rank-1 and mAP were further increased to 88.40% and 90.61% respectively. The method proposed in this paper has achieved significant performance improvement on the RegDB dataset. In the VIS→IR mode, Rank-1, Rank-10 and mAP reached 95.54%, 97.96% and 92.10% respectively; In the IR→VIS mode, the corresponding indicators are 95.95%, 96.45% and 92.79%. These data fully prove the superior performance of this method in cross-modal matching, anti-background noise and other aspects. Especially when dealing with large-scale and complex environmental data, it shows strong practicability and adaptability, and can effectively deal with various challenges in actual monitoring scenarios.

Table 1: Comparison with state-of-the-art on SYSU-MM01 of Rank-k Accuracies (%) and mAP (%)

Methods	All Search			Indoor Search		
	Rank=1	Rank=10	mAP	Rank=1	Rank=10	mAP
Hi-CMD[21]	34.9%	77.6%	35.9%	—	—	—
NFS[22]	56.9%	91.3%	55.5%	62.8%	96.5%	69.8%
MCLNet[23]	65.8%	93.3%	62.0%	72.6%	97.0%	76.6%
DART[24]	68.7%	96.4%	66.3%	72.5%	97.8%	78.2%
MPANet[25]	70.6%	96.2%	68.2%	76.7%	98.2%	81.0%
DEEN[26]	74.7%	97.6%	71.8%	80.3%	99.0%	83.3%
MUN[27]	76.24%	97.84%	73.81%	79.42%	98.09%	82.06%
IDKL[28]	81.42%	97.38%	79.85%	87.14%	98.28%	89.37%
D-selacfp	82.77%	98.31%	79.91%	88.40%	98.82%	90.61%

Table 2: Comparison with state-of-the-art on RegDB of Rank-k Accuracies (%) and mAP (%)

Methods	Visible-Infrared			Infrared-Visible		
	Rank=1	Rank=10	mAP	Rank=1	Rank=10	mAP
Hi-CMD[21]	70.9%	86.4%	66.0%	-	-	-
NFS[22]	80.5%	91.6%	72.1%	78.0%	90.5%	69.8%

MCLNet[23]	80.3%	92.7%	73.1%	75.9%	90.9%	69.5%
DART[24]	83.6%	-	75.7%	82.0%	-	73.8%
MPANet[25]	82.8%	-	80.7%	83.7%	-	80.9%
DEEN[26]	91.1%	97.8%	85.1%	89.5%	96.8%	83.4%
MUN[27]	95.19%	-	87.15%	91.86%	-	85.01%
IDKL[28]	94.72%	-	90.19%	94.22%	-	90.43%
D-selacfp	95.54%	97.96%	92.10%	95.95%	96.45%	92.79%

4. CONCLUSION

This paper proposes D-SELACFP: a cross-modal person re-identification network incorporating Local Attention Module (LAM) and Inter-modal Feature Perception Enhancement (I-MFPE). The framework effectively integrates multi-granularity features through a non-local fusion architecture while enhancing pedestrian feature representation via cross-scale perception strategies. To mitigate modality discrepancies, LAM suppresses noise interference and amplifies discriminative information through frequency-domain feature compression that selectively enhances identity-critical components. The I-MFPE module further optimizes feature extraction through its cross-scale guidance mechanism. Experimental validation on SYSU-MM01 and RegDB datasets confirms the method's efficacy in reducing modality gaps and improving recognition accuracy. Visual results demonstrate its capability to maintain robust performance despite complex backgrounds. Future research will focus on refined feature extraction algorithms and unsupervised learning strategies to enhance real-world generalization capabilities.

REFERENCES

[1] LUO H, JIANG W, FAN X, et al. A survey on deep learning based per- son re-identification[J]. Acta Automatica Sinica, 2019, 45(11): 2032- 2049.
[2] LIAO S, SHAO L. Graph sampling based deep metric learning for generalizable person re - identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7359-7368.
[3] 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.
[4] HAo Y, WAnG N N, GAo X B, et al. Dual-alignment feature embedding for cross-modality person re-identification [C]//27th ACM International Conference on Multimedia, 2019: 57-65.
[5] CHENG D, Li X H, Qi M B, et al. Exploring cross-modality commonalities via dual-stream multi-branch network for infrared-visible person re-identification [J]. IEEE Access, 2020, 8: 12824-12834.
[6] Liu H J, ChEnG J, WAnG W, et al. Enhancing the discriminative feature learning for visiblethermal cross-modality person re-identification [J]. Neurocomputing, 2020, 398: 11-19.
[7] WANG P Y, ZhAo Z C, Su F, et al. Deep multi-patch matching network for visible thermal person re-identification [J]. IEEE Transactions on Multimedia, 2021, 23: 1474-1488.
[8] Zhu Y X, YAnG Z, WAnG L, et al. Hetero-center loss for cross-modality person reidentification [J]. Neurocomputing, 2020, 386: 97-109.
[9] Fu C Y, Hu Y B, Wu X, et al. CM-NAS: cross-modality neural architecture search for visibleinfrared person re-identification [C]//IEEE/CVF International Conference on Computer Vision, 2021: 11803-11812.
[10] YE M, RuAn W J, Du B, et al. Channel augmented joint learning for Visible-Infrared recognition [C]//IEEE/CVF International Conference on Computer Vision, 2021: 13547-13556.
[11] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[12] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
[13] Yang J, Zhang J, Yu F, et al. Learning to know where to see: a visibility-aware approach for occluded person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11885-11894.
[14] Liu Z, Zhang Z, Li D, et al. Dual-branch self-attention network for pedestrian attribute recognition[J]. Pattern Recognition Letters, 2022, 163: 112-120.
[15] Li C, Yang X, Yin K, et al. Pedestrian re-identification based on attribute mining and reasoning[J]. IET Image Processing, 2021, 15(11):2399-2411.

[16] Zhao J, Wang H, Zhou Y, et al. Spatial-channel enhanced transformer for Visible-Infrared person reidentification[J]. IEEE Transactions on Multimedia, 2022, 25: 3668-3680.
[17] QIN Z, ZHANG P, WU F, et al. Fcanet: Frequency channel attention networks[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 783-792
[18] YANG K W, YANG J W, TIAN X M. Learning multi-granularity fea tures from multi-granularity regions for person re-identification[J]. Neurocomputing, 2021, 432: 206-215.
[19] LI Q, WANG H J, LI B Y, et al. Fast image semantic segmentation method based on improved IIE-SegNet [J]. Journal of Harbin Engineer ing University, 2024, 45(2): 314-323.
[20] WU A C, ZHENG W S, YU H X, et al. RGB-infrared cross-modality person re-identification [C]//Proceedings of the 2017 IEEE Interna tional Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 5390-5399.
[21] CHOI S, LEE S M, KIM Y, et al. Hi-CMD: hierarchical cross-modality disentanglement for Visible-Infrared person re-identification[C]//Pro- ceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 10257- 10266.
[22] CHEN Y, WAN L, LI Z H, et al. Neural feature search for RGBinfrared person re-identification [C]//Proceedings of the 2021 IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 587-597.
[23] HAO X, ZHAO S Y, YE M, et al. Cross-modality person reidentification via modality confusion and center aggregation[C]//Pro- ceedings of the 2021 IEEE/CVF International Conference on Com- puter Vision (ICCV). Piscataway: IEEE Press, 2021: 16403-16412.
[24] YANG M X, HUANG Z Y, HU P, et al. Learning with twin noisy la- bels for Visible-Infrared person re-identification[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recogni- tion (CVPR). Piscataway: IEEE Press, 2022: 14288-14297.
[25] WU Q, DAI P Y, CHEN J, et al. Discover cross-modality nuances for Visible-Infrared person re-identification[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 4328-4337.
[26] PARK H, LEE S, LEE J, et al. Learning by aligning: Visible-Infrared person re-identification using cross-modal correspondences[C]//Pro- ceedings of the 2021 IEEE/CVF International Conference on Com- puter Vision (ICCV). Piscataway: IEEE Press, 2021: 12026-12035.
[27] YU H, CHENG X, PENG W, et al. Modality unifying network for Visible-Infrared person re-identification [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 11185-11195.
[28] REN K, ZHANG L. Implicit Discriminative Knowledge Learning for Visible-Infrared Person Re-Identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 393-402.