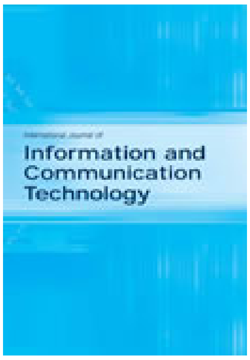


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Multi-modal similarity feature exchange and structural perception for person re-identification

Xuefeng Lei

DOI: 10.1504/IJICT.2025.10074510

Article History:

Received:	08 July 2025
Last revised:	18 August 2025
Accepted:	24 August 2025
Published online:	20 November 2025

Multi-modal similarity feature exchange and structural perception for person re-identification

Xuefeng Lei

College of Artificial Intelligence,
Jiangxi Industry Polytechnic College,
Nanchang, Jiangxi, 330095, China
Email: leixuefeng_123456@163.com

Abstract: Visible-infrared person re-identification is crucial for surveillance, aiming to match person images across visible and infrared modalities. However, spectral and style gaps hinder local structure modelling and cross-modal feature alignment. We propose the cross-modality similarity exchange transformer (CSET) to improve both aspects. CSET uses two modality-specific transformer encoders to extract features independently. A similarity exchange mechanism computes intra-modality similarity and cross-modality Jaccard distance, selectively exchanging correlated token features for local alignment and feature complementation. To enhance structural perception, we introduce a multi-relational heterogeneous graph attention mechanism, building a graph from transformer outputs where positional embedding differences define relation levels. Feature aggregation is guided by relational strength to capture fine-grained structural cues. Experiments on RegDB and SYSU-MM01 show CSET outperforms state-of-the-art methods in Rank-1 accuracy and mAP, validating its cross-modal learning effectiveness.

Keywords: cross-modality person re-identification; similarity exchange; SE; transformer; heterogeneous graph attention; feature alignment.

Reference to this paper should be made as follows: Lei, X. (2025) ‘Multi-modal similarity feature exchange and structural perception for person re-identification’, *Int. J. Information and Communication Technology*, Vol. 26, No. 41, pp.24–42.

Biographical notes: Xuefeng Lei is an Associate Professor, received his Master’s and graduated from East China Normal University in 2010. He worked in Jiangxi Industry Polytechnic College. His research interests include mobile application development and artificial intelligence.

1 Introduction

Person with visible-infrared re-identification (VI-ReID), which seeks to match the same individual across many modalities, is extensively used in domains including smart cities, security protection, and nocturnal surveillance (Feng et al., 2023; Yu et al., 2023; Chen et al., 2023a). There is a notable feature distribution disparity between visible and infrared pictures because visible photos are susceptible to noise in different lighting conditions while infrared images lack rich colour and texture information. This discrepancy prevents traditional single-modality person re-identification methods from being directly applicable to VI-ReID tasks (Liu et al., 2024; Lu et al., 2023). Therefore, how to effectively align feature distributions across modalities and achieve robust consistent representations has become a core challenge in cross-modality person re-identification research (Kim et al., 2023; Chong, 2023).

To mitigate the impact of modality discrepancies on retrieval performance, existing studies mainly explore two directions: feature alignment and modality-invariant feature extraction (Fang et al., 2023; Wu and Ye, 2023; Chang et al., 2024). One category of methods employs parameter-sharing single-stream networks combined with adversarial learning techniques to guide the mapping of visible and infrared images into a unified feature space, thereby alleviating distribution discrepancies between modalities (Wei et al., 2023). However, these methods often overly rely on adversarial loss, leading to insufficient modality feature fusion and limited capture of local fine-grained information (Zhang et al., 2023). Another category of methods adopts dual-stream structures, extracting visible and infrared features separately and narrowing modality gaps through local or global alignment mechanisms. Although these approaches improve feature consistency to some extent, they tend to introduce feature shifts and still have limited capability in modelling local structural details (Chen et al., 2023b). In addition, some methods utilise a staged training strategy, where single-modality features are pre-trained before cross-modality alignment. While this improves overall consistency, the feature interaction process remains relatively simple, and local structural relationships are not fully exploited (Huang et al., 2023; Shi et al., 2024). In recent years, transformer architectures, owing to their excellent self-attention modelling capabilities, have been introduced into cross-modality person re-identification tasks (Zhang and Wang, 2023; Pang et al., 2023). Through adaptive learning, different modality features can achieve a certain degree of fusion and alignment. Nevertheless, existing transformer-based methods generally face two critical issues. On the one hand, feature interaction across modalities typically focuses on high-level semantic features, neglecting the modelling of low-level local structure consistency (Shi et al., 2023; Sarker and Zhao, 2024). On the other hand, when modelling relationships among tokens, transformers usually assume uniform connection strength among all nodes, failing to effectively distinguish actual relational strengths, which limits the model's ability to highlight key local structures or suppress irrelevant regions during feature aggregation, thereby restricting the perception and modelling capabilities of complex cross-modality structural information (Pan et al., 2024). Thus, current methods still exhibit significant limitations in modality fusion and fine-grained structural alignment. There is an urgent need to design new mechanisms to strengthen consistent modelling of local and global features across modalities and further enhance the accuracy and robustness of cross-modality person retrieval.

To address the aforementioned issues, we propose the CSET to enhance the local alignment and structural perception capabilities of cross-modality features. To promote consistent alignment of local features across modalities, we introduce the similarity exchange (SE) mechanism during the transformer encoding process. Specifically, after each transformer layer output, we compute the similarity matrices among tokens within visible and infrared images respectively to capture intra-modality local structural relationships. Meanwhile, we compute the Jaccard distance between cross-modality tokens to measure feature similarity across modalities. Based on the intra and cross-modality similarity information, the SE mechanism dynamically selects the most correlated cross-modality token pairs and performs feature exchanges at specific positions, thereby enhancing feature interaction and fusion at the local level. Through layer-by-layer feature exchange, the model can continuously narrow modality discrepancies during feature extraction and improve the consistency of local structures. Moreover, to further enhance the structural modelling capabilities of cross-modality features, CSET introduces a multi-relational heterogeneous graph attention (MHGA) mechanism based on the deep transformer features. This mechanism treats transformer output tokens as graph nodes, and, combined with positional embedding information, categorises node relationships into three levels based on positional relations: same position, neighbouring position, and distant position. Different weight coefficients are assigned according to node relationships to construct a relation-aware heterogeneous graph structure. Subsequently, node features are aggregated through a graph attention mechanism, enabling the model to differentiate relational strengths among nodes during feature aggregation and effectively model local neighbourhood and global structural features. Finally, by differentially modelling various relational levels, the model enhances the spatial structural representation capabilities of cross-modality features.

In summary, our main innovations and contributions are as follows:

1 We propose the SE mechanism. By dynamically performing local feature interaction at each transformer layer, and guiding local feature consistency alignment through the fusion of intra-modality token similarity and cross-modality Jaccard distance, we effectively alleviate modality distribution discrepancies.

2 We propose the MHGA mechanism. By constructing a heterogeneous graph based on positional relationships and employing a hierarchical attention mechanism to model different node relationships at a fine-grained level, we enhance the structural perception and expressive capabilities of cross-modality features.

3 We propose the CSET model, which effectively addresses the shortcomings of existing VI-ReID methods in local alignment and structural modelling. It significantly improves cross-modality person re-identification performance on multiple mainstream datasets, fully validating the effectiveness and application potential of our method.

2 Related work

2.1 Feature alignment in cross-modal person re-identification

One of the main issues with cross-modal person re-identification tasks is feature alignment. Research has mostly focused on using rich semantic information to help align in order to minimise feature disparities between modalities. Zhai et al. (2024) proposed to generate fine-grained attribute descriptions and incorporate various prompt information to enhance the semantic richness of image features. However, this method relies on externally generated language information, which may be influenced by subjectivity and bias in practical applications, and it struggles to fully capture the internal local structural relationships within images. Building on this, Yang et al. (2025) introduced large-scale pre-trained vision-language models into cross-modal retrieval, optimising the mapping of visual and textual features into a shared space through contrastive learning. Although this method effectively improves global semantic consistency, it still shows limitations in local feature alignment, particularly in filtering irrelevant information and avoiding redundant interference.

To further enhance local structural feature modelling, Huang et al. (2024) proposed a multi-scale dynamic feature alignment mechanism that combines enhanced textual global perception modules with human structure association modelling to improve the consistency and discriminability of cross-modal features. Although this approach enhances global representation capabilities, its local alignment performance remains limited in scenes with strong subjectivity or drastic scale variations. To address the limitations in fine-grained fusion, Wu et al. (2024a) designed an implicit learning transformer framework that excavates cross-modal implicit relationships through a bidirectional masking mechanism and introduces a cross-modal similarity matching module to optimise feature consistency. Despite overall performance improvements, due to the lack of explicit local alignment modelling, the fine-grained feature fusion ability under complex scenarios remains insufficient. Furthermore, from the perspective of multi-level feature alignment, Li et al. (2024a) proposed a cascaded alignment framework that progressively reduces modality distribution differences at the input, frequency, and local part levels. Although hierarchical alignment achieved good results, the overall framework is complex, training is challenging, and its adaptability to dynamic or complex scenes still has room for improvement.

2.2 Transformer in cross-modal person re-identification

In recent years, transformer topologies have been frequently used in cross-modal person re-identification tasks because of their superior long-range dependency modelling capabilities. To effectively exploit the complementarity among multi-modal features, Zheng et al. (2024) proposed a transformer-based relational regularisation method, achieving feature fusion and optimisation across modalities through adaptive collaborative matching and embedding enhancement modules. However, this method still handles local structural features rather coarsely and lacks fine-grained modelling of regional feature relationships.

Subsequently, Sarker et al. (2024) conducted a systematic review of transformer-based person re-identification research, summarising the advantages of transformers in feature extraction and cross-modal matching, while also pointing out that current methods still lack robustness and generalisation in handling appearance changes, occlusions, and scale variations under complex environments, requiring further enhancement of the adaptability of transformer structures. In the direction of video-level cross-modal person re-identification, Feng et al. (2024) proposed a cross-modal spatiotemporal transformer that models local spatiotemporal block information in a pipelined manner and introduces a multi-frame fusion module to capture long-term temporal dependencies. Although this method effectively extends the modelling range beyond traditional convolutional networks, its modelling capability for static local details remains limited when directly applied to static image scenarios.

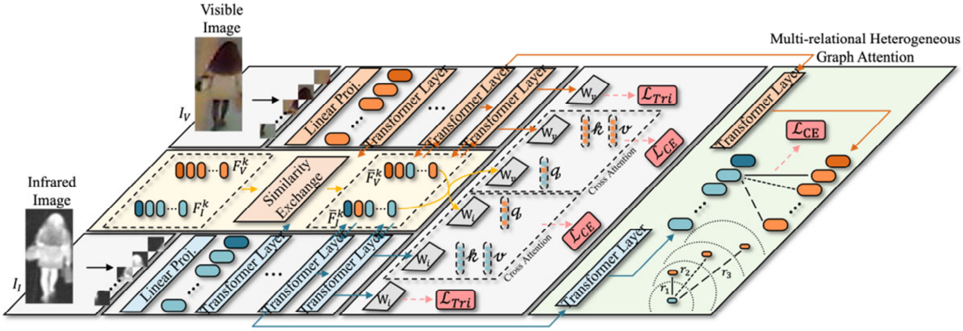
To better incorporate local fine-grained information, Li et al. (2024b) proposed a multi-granularity cross-modal transformer network, designing a pyramid partitioning and cross-layer feature interaction mechanism to progressively mine salient features from local to global levels. Although this method improves the utilisation of local information, its model structure is relatively complex, and there are certain inference efficiency bottlenecks in large-scale retrieval scenarios. Additionally, to address the representation difficulties caused by modality discrepancies, An et al. (2024) proposed a new framework based on hybrid data augmentation and transformer feature extraction, introducing context broadcasting and modality-shared centre loss to enhance the consistency of cross-modal features. Although modality distribution differences are alleviated to some extent, the robustness of the method under small-sample and extreme conditions in real-world environments still needs improvement.

Most existing methods focus on feature space alignment or rely on transformer modelling of global features but often neglect consistent modelling of local fine-grained structures and differentiated modelling of inter-node relationships, resulting in insufficient modality fusion and weak structural perception capabilities. In contrast, our proposed method not only dynamically mines and exchanges the most correlated cross-modal features within transformer layers to explicitly promote local structure alignment, but also introduces multi-relational heterogeneous graph modelling at deep feature levels to finely distinguish the association strengths between different nodes.

3 Proposed method

We suggest the CSET as a solution to the problems caused by the notable modality discrepancies between visible and infrared pictures as well as the difficulties in cross-modal feature alignment. Figure 1 depicts the CSET pipeline as a whole.

Figure 1 The pipeline of the CSET model (see online version for colours)

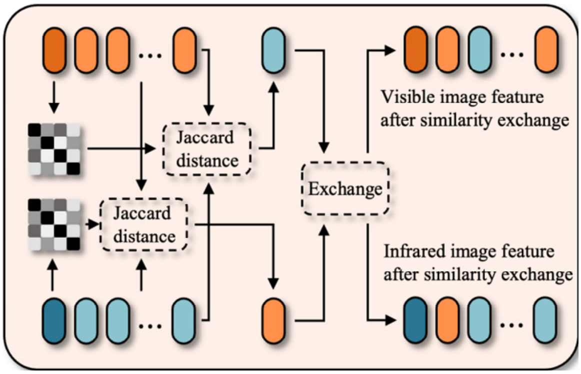


Specifically, the objective is to learn a unified cross-modal feature representation for a visible image I_V and an infrared image I_I. CSET employs two transformers with non-shared parameters to process visible and infrared images separately. First, feature tokens, a class token, and positional embeddings are obtained through linear projection. Then, deep features are extracted through L transformer layers. Between the k^{th} and $(k + 1)^{\text{th}}$ transformer layers, we design a SE mechanism to promote cross-modal feature fusion. The intra-modality token similarity matrices A_V and A_I are computed separately from the visible features F_V^k and infrared features F_I^k. Subsequently, by jointly considering the Jaccard distance and cross-modal token similarities, the most similar tokens are selected from the opposite modality for feature exchange, resulting in the exchanged and enhanced features $\overline{F}_V^k$ and $\overline{F}_I^k$. The output features F_V^L and F_I^L from the final transformer layer are respectively fed into feedforward networks W_v, W_v' and W_i, W_i'. Additionally, the cross-modal features after SE are also input into the corresponding feedforward networks. Among them, the output features from W_v' and W_i' are used to compute the triplet loss, while the outputs from W_v and W_i are used to generate mixed features q and clean features k, v, followed by cross-attention computation and cross-distillation loss calculation. Moreover, the features $F_V^{(L-1)}$ and $F_I^{(L-1)}$ extracted from the (L–1)th transformer layer are fed into another transformer layer with independent parameters, where we design the MHGA mechanism. In this mechanism, tokens are treated as graph nodes, and three levels of node relationships are constructed based on positional embeddings. Feature aggregation is performed through relation-aware attention, further optimising cross-modal feature representation. Finally, the concatenated features are used to compute the cross-entropy loss, supervising the entire network training.

3.1 Similarity exchange mechanism

We provide the SE method to facilitate token exchange across modalities in order to efficiently close the feature gap between various modalities in the cross-modal person re-identification problem. The general procedure is depicted in Figure 2.

Figure 2 The pipeline of the similarity exchange mechanism (see online version for colours)


Jaccard distance
Jaccard distance
Exchange
Visible image feature after similarity exchange
Infrared image feature after similarity exchange

Specifically, let the output visible features of the kth transformer layer be denoted as $F_V^k \in \mathbb{R}^{N \times d}$ and the infrared features as $F_I^k \in \mathbb{R}^{N \times d}$, where N represents the number of tokens and d denotes the feature dimension of each token. First, we calculate the intra-modality token similarity matrices separately to capture the structural relationships among tokens within each modality, as shown in equation (1):

$$A_V = Softmax\left(F_V^k \left(F_V^k\right)^\top\right), A_I = Softmax\left(F_I^k \left(F_I^k\right)^\top\right) \tag{1}$$

where A_V, $A_I \in \mathbb{R}^{N \times N}$, and each element $A_{V,ij}$ or $A_{I,ij}$ represents the similarity between the ith and jth tokens. Subsequently, to measure the structural differences between cross-modal tokens, we introduce the Jaccard distance to assess the similarity between two tokens represented as vectors. Specifically, for the ith token $F_{I,i}^k$ in the infrared features and the j^{th} token $F_{V,j}^{k}$ in the visible feature set F_{V}^{k}, the Jaccard distance is defined as in equation (2):

$$J_{(I\to V),i,j} = 1 - \frac{F_{I,i}^{k} \cdot F_{V,j}^{k}}{\left\| F_{I,i}^{k} \right\|^{2} + \left\| F_{V,j}^{k} \right\|^{2} - F_{I,i}^{k} \cdot F_{V,j}^{k}} \tag{2}$$

where ||·|| denotes the vector norm, and $J_{(I\to V),i,j}$ indicates the structural difference between the i^{th} infrared token and the j^{th} visible token. A smaller distance value implies greater structural similarity between the two tokens. Next, to simultaneously consider the local structure within each modality and the structural differences across modalities, we design a mapping function $M(\cdot)$ to integrate the intra-modality token similarity matrix and the cross-modality Jaccard distance matrix, thus obtaining a cross-modality matching weight matrix, as shown in equation (3):

$$S_{I\to V} = M\left(A_I, J_{(I\to V)}\right) = \alpha \cdot \left(1 - J_{(I\to V)}\right) + (1-\alpha) \cdot A_I \tag{3}$$

where α is a balance coefficient ranging between [0, 1], controlling the contribution between local similarity and cross-modal similarity.

3.2 Multi-relational heterogeneous graph attention mechanism

To further capture richer and finer structural information between cross-modal features and effectively distinguish the contribution degrees of different node relationships, we propose the MHGA mechanism. Specifically, the outputs from the $(L-1)^{\text{th}}$ transformer layer are fed into another transformer layer with independent parameters. The output visible features are denoted as $F_V^L \in \mathbb{R}^{N\times d}$ and the infrared features as $F_I^L \in \mathbb{R}^{N\times d}$. These two modalities' features are unified into a node set $V = \{v_1, v_2, \ldots, v_{2N}\}$, where each node v_i has an initial feature representation $h_i \in \mathbb{R}^d$. To more accurately characterise the positional structural relationships between token nodes, we explicitly divide the relationships between nodes into three levels. Different relational strength weights are assigned based on the positional embedding differences between nodes, forming a relational strength matrix $R \in \mathbb{R}^{2N\times 2N}$, defined as in equation (4):

$$R_{i,j} = \begin{cases} \lambda_1, & \textit{if } pos(i) = pos(j) \\ \lambda_2, & \textit{if } \left|pos(i) - pos(j)\right| = 1 \\ \lambda_3, & \textit{otherwise} \end{cases} \tag{4}$$

where nodes i and j originate from different tokens, pos(i) denotes the positional index of node v_i in the feature sequence, and $\lambda_1 > \lambda_2 > \lambda_3 > 0$ represent gradually decreasing relational strengths: the strongest relationship exists between nodes at the same position, followed by adjacent nodes, and the weakest between distant nodes. Subsequently, by combining the feature similarity and the positional relational strength between node pairs, the raw attention weight between node i and node j is defined as in equation (5):

$$e_{ij} = Leaky\ ReLU\left(a^{\top}\left[Wh_i \,\|\, Wh_j\right]\right) \times R_{i,j} \tag{5}$$

where $W \in \mathbb{R}^{d' \times d}$ is a shared linear transformation matrix, $a \in \mathbb{R}^{2d'}$ is a learnable attention vector, and the symbol // denotes vector concatenation. By introducing $R_{i,j}$, the relational strength between different levels of nodes differentially contributes to the attention mechanism. Then, the attention weights over all neighbouring nodes N_i of node i are normalised using the Softmax function, as shown in equation (6):

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})} \tag{6}$$

where α_{ij} represents the importance weight of node j when aggregating features to node i, and N_i denotes the set of all neighbour nodes connected to node i. Furthermore, the final updated feature h_i' for node i can be obtained by the weighted aggregation of the neighbour nodes' features, as shown in equation (7):

$$h_i' = \sum_{j \in N_i} \alpha_{ij} W h_j \tag{7}$$

Through the above differentiation process, the MHGA mechanism can distinguish contributions from different relational levels, thereby enhancing the structural perception capability between cross-modal node features.

3.3 Discussion

In the SE mechanism, to measure the structural similarity between cross-modal tokens, we adopt the Jaccard distance instead of conventional metrics such as Euclidean distance, cosine distance, or other similarity measures. We discuss the rationale behind this choice as follows.

Firstly, the Jaccard distance effectively measures the degree of intersection between feature subspaces, emphasising the overlap proportion of jointly activated regions between feature vectors. This characteristic is particularly important for cross-modal person re-identification because visible and infrared images naturally differ in spectral characteristics and visual styles. Directly using Euclidean or Cosine distances is prone to being affected by overall feature scale variations or modality shifts, leading to inaccurate cross-modal matching. In contrast, the Jaccard distance, by focusing on the commonality between features rather than absolute differences, can more robustly capture modality-independent structural similarities and enhance the consistency of local feature alignment.

Secondly, the Jaccard distance has a natural advantage when dealing with sparse feature distributions or localised response features. Since token features extracted by the transformer are often sparse across different regions, with only a few tokens carrying major discriminative information, the Jaccard distance can naturally ignore irrelevant feature regions and concentrate on the truly intersecting structural information. This helps avoid noise interference and improves the stability and robustness of the SE mechanism under complex scenarios.

3.4 Loss function

We use two loss functions, cross-entropy loss and triplet loss, to efficiently direct cross-modal feature learning while concurrently optimising feature discriminability and modality alignment. While the triplet loss is used to optimise the relative distance relationships in the feature space by pushing apart samples of different identities and pulling closer samples of the same identity across different modalities, the cross-entropy loss mainly oversees the classification task to guarantee that features belonging to the same identity are correctly classified.

First, the cross-entropy loss is used in the classification process. Let the network output the predicted class probability distribution $\hat{y}$ and the ground-truth label be y. The cross-entropy loss is defined as shown in equation (8):

$$L_{CE} = -\sum_{c=1}^{C} y_c \log\left(\hat{y}_c\right) \tag{8}$$

where C denotes the total number of classes, y_c is the ground-truth indicator for class c, and $\hat{y}_c$ is the predicted probability of belonging to class c. By minimising L_{CE}, the separability of features can be effectively enhanced, enabling the network to produce features with better class discriminability. Secondly, to further improve the consistency of the cross-modal feature space, we also adopt the triplet loss. Let f_a, f_p, and f_n represent the feature embeddings of the anchor, positive, and negative samples, respectively. The triplet loss is defined as in equation (9):

$$L_{Tri} = \sum_{i=1}^{N} \left[\left\| f_a^i - f_p^i \right\|_2^2 - \left\| f_a^i - f_n^i \right\|_2^2 + m \right]_+ \tag{9}$$

where N denotes the number of sample triplets, $\|\cdot\|_2$ represents the Euclidean distance, m is a preset margin hyperparameter (set to 0.3 in this work), and $[\cdot]_+$ denotes the positive part operator. By minimising L_{Tri}, the network is encouraged to pull features of the same identity closer together in the feature space while pushing features of different identities farther apart, thus enhancing the robustness of cross-modal retrieval. Finally, the total loss function in our method is a weighted combination of the cross-entropy loss and the triplet loss, defined as in equation (10):

$$L_{total} = L_{CE} + \beta L_{Tri} \tag{10}$$

where β is a balancing coefficient (set to 0.7 in this work) to control the relative contributions of the cross-entropy loss and the triplet loss to the final training objective.

4 Experiments

4.1 Experimental settings and environment

To validate the effectiveness of the proposed CSET model, we conduct experimental evaluations under standard hardware and software frameworks. All experiments are carried out on servers running the Ubuntu 20.04 operating system in terms of the hardware environment. An NVIDIA RTX 3090 24GB GPU, 64 GB of RAM, and an Intel

Xeon Gold 6226R CPU make up the system setup. In terms of the software environment, PyTorch 1.12 is utilised as the deep learning framework, while Python 3.8 is chosen as the main programming language. For hyperparameter settings, the initial learning rate is set to 3 × 10^{-4}, and the Adam optimiser is employed for parameter updates. The weight decay is set to 5 × 10^{-4} to prevent overfitting. The batch size is set to 64, and the maximum number of training epochs is 120. The dropout rate is set to 0.5 to further enhance the model's generalisation ability. The input images are uniformly resized to 256 × 128. In the SE mechanism, the parameter α, which balances intra-modality local similarity and cross-modality similarity, is set to 0.7. In the MHGA mechanism, the strength parameters for different levels of node relationships are set as $\lambda_1 = 1$, $\lambda_2 = 0.7$ and $\lambda_3 = 0.3$, ensuring that the relationship strength decreases gradually with increasing distance between nodes. Additionally, the learning rate is adjusted during training using the StepLR strategy, decaying to 0.1 times the original learning rate every 40 epochs to accelerate model convergence.

4.2 Datasets and evaluation metrics

We conduct experiments on two mainstream visible-infrared (VI) cross-modal person re-identification datasets: SYSU-MM01 (Wu et al., 2017) and RegDB (Nguyen et al., 2017).

The SYSU-MM01 collection covers 491 distinct pedestrian IDs with 286,628 visible photos and 15,792 infrared images. Photos from the remaining 96 identities are included in the testing set, while 22,258 visible and 11,909 infrared photos from 395 identities make up the training set. 301 visible photographs are chosen at random from the remaining test set to create the gallery set, while 3,803 infrared images are utilised as the query set. Furthermore, SYSU-MM01 has two testing modes: indoor-search mode, which limits retrieval to indoor photos only, and all-search mode, which uses all test images.

Each of the 412 pedestrian IDs in the RegDB collection has ten visible and ten infrared photos. We choose 206 identities at random for training and the remaining 206 identities for testing in accordance with the conventional split technique. Ten distinct random splits of the RegDB dataset are carried out, and each split is subjected to independent training and testing in accordance with standard procedure in order to provide consistent and trustworthy experimental findings.

We use mean average precision (mAP) and cumulative matching characteristics (CMC) curves as the main assessment measures in order to thoroughly assess the models' performance in cross-modal person re-identification tasks. The mAP thoroughly takes into account both the accuracy and recall of the retrieval results, whereas the CMC curve calculates the correct matching rate at various retrieval ranks.

4.3 Results and analysis

4.3.1 Ablation study

We plan a rigorous ablation investigation to confirm the precise contributions of each important module to the overall performance. We assess the impacts of the MHGA, cross attention (CA), and SE mechanisms in the model one after the other. We conduct extensive experiments on the RegDB dataset, summarise the results in Table 1, and discuss the relationship between model accuracy and complexity in Table 2.

Table 1 The impact of different mechanisms on the performance of the CSET model

SE			*CA*	*MHGA*	*RegDB*			
					I2V		*V2I*	
I2V	*V2I*	*I2V&V2I*			*R1*	*mAP*	*R1*	*mAP*
					89.63	84.12	90.32	84.35
√					91.96	86.20	90.44	84.49
	√				89.76	84.32	92.72	86.57
		√			92.49	87.73	93.75	88.32
		√	√		93.33	88.87	94.36	89.35
		√	√	√	94.16	89.72	94.93	90.26

As shown in Table 1, the baseline achieves Rank-1 accuracies of 89.63% (I2V) and 90.32% (V2I). Introducing the SE mechanism in only one direction significantly boosts performance for that specific direction – 91.96% for I2V and 92.72% for V2I – while having little effect on the other. This highlights SE's targeted enhancement of local feature consistency.

When applied bidirectionally, SE further improves both directions, achieving 92.49% (I2V) and 93.75% (V2I), along with higher mAP scores. These results confirm that bidirectional exchange promotes more effective local feature interaction, reduces modality gaps, and improves overall retrieval performance.

On the basis of bidirectional exchange, the further introduction of the CA mechanism leads to additional performance gains, with the I2V and V2I Rank-1 accuracies improving to 93.33% and 94.36%, respectively. The CA mechanism, by introducing interactions between the mixed feature q and the clean features k and v, enables deeper fusion of different modality features in the highlevel feature space, effectively compressing modality differences and enhancing discriminative capability in local region matching.

Finally, when the MHGA mechanism is introduced, the model achieves Rank-1 accuracies of 94.16% and 94.93% in the I2V and V2I retrieval directions, respectively, with corresponding increases in mAP to 89.72% and 90.26%. By modelling multi-level relationships among nodes, MHGA effectively distinguishes feature associations at different positions and across modalities at a fine-grained level, thereby enhancing the structural perception and robustness of overall feature representations.

Table 2 The impact of different modules on model complexity

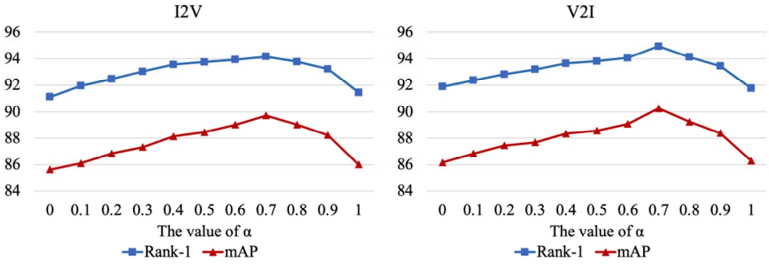
SE			*CA*	*MHGA*	*Parameters (M)*	*inference time(ms)*
I2V	V2I	I2V&V2I				
					22.8	5.4
√					23.5	5.7
	√				25.1	6.2
		√			25.8	6.5
		√	√		24.3	6.0
		√	√	√	26.6	6.9

From Table 2, it can be seen that the SE, CA, and MHGA modules, while improving performance, introduce only a relatively limited increase in model complexity. The SE module adds only a small amount of parameters and inference overhead, the CA module incurs slightly higher computational cost than SE, and the MHGA module, due to its multi-relational graph construction and multi-head attention computation, results in a more noticeable increase in parameters and inference time. When all three modules are used together, the parameters and inference time increase only slightly compared to the baseline, indicating that the proposed method maintains high accuracy while still offering good computational efficiency and deployment potential.

4.3.2 Hyperparameter experiments

To explore the importance of the fusion between intra-modality local structure and cross-modality similarity in the SE mechanism, we conduct a hyperparameter sensitivity study on the balance coefficient α in the mapping function. The coefficient α controls the weighting between the intramodality similarity matrix and the cross-modality Jaccard distance. Experiments are conducted on the RegDB dataset with different values of α, and the results are shown in Figure 3.

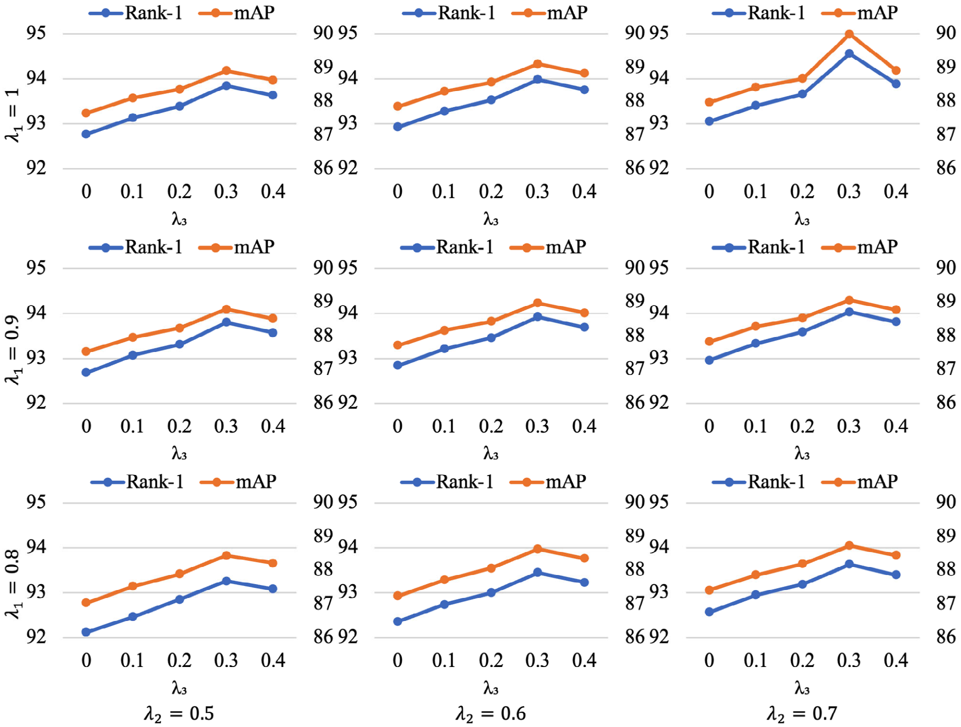
Figure 3 The impact of the balance coefficient on model performance (see online version for colours)



As shown in Figure 3, as the balance coefficient α gradually increases, the performance of the model in both infrared-to-visible (I2V) and visible-to-infrared (V2I) retrieval directions shows an overall upward trend. At the initial stage, when α is small, the model mainly relies on intra-modality local structural information, resulting in relatively low retrieval accuracy. As α increases, the model gradually incorporates more crossmodality structural difference information, effectively enhancing the consistency and discriminability of features and steadily improving retrieval performance. When α is set to 0.7, the model achieves peak performance in both retrieval directions, indicating that the fusion between intra-modality local structure and cross-modality similarity reaches an optimal balance, significantly improving the effectiveness of cross-modal feature alignment. However, when α is further increased to 0.8 and beyond, retrieval performance declines, suggesting that excessive reliance on cross-modality similarity while neglecting local structural relationships can impair the integrity and stability of feature representations, ultimately leading to a decrease in overall performance.

To further investigate the effect of the node relationship strength parameters in the MHGA mechanism on cross-modal feature modelling, we conduct an analysis of λ_1, λ_2, and λ_3. The experimental results are shown in Figure 4.

Figure 4 The impact of different , and values on model performance (see online version for colours)



Overall, regardless of the specific λ_1 and λ_2 settings, both Rank 1 accuracy and mAP show consistent variation trends with changes in λ_3. When λ_3 is small, the model performance is relatively low; as λ_3 gradually increases, performance steadily improves, reaching a peak when $\lambda_3 = 0.3$, followed by a slight decline.

Specifically, when λ_3 is small, the feature interaction effect among distant nodes is weak, leading to insufficient global structural modelling capability. The model mainly relies on local neighbourhood information, resulting in weaker consistency and robustness of cross-modal features, and thus lower Rank-1 and mAP scores. As λ_3 increases, the model's perception of distant node features gradually strengthens, enabling more comprehensive capture of complex structural relationships between modalities, significantly enhancing matching accuracy and overall retrieval precision. When λ_3 reaches 0.3 , both Rank-1 accuracy and mAP achieve their optimal values, indicating that a good trade-off between local and global relationship modelling is achieved, balancing fine-grained feature representation and large-scale feature association. However, when λ_3 is further increased to 0.4, performance slightly declines. This is because overly high distant node weights weaken the importance of local structures, introduce more irrelevant or noisy features, degrade feature representation ability, weaken local finegrained discriminability, and ultimately affect overall retrieval performance.

Additionally, different settings of λ_1 and λ_2 also have a certain impact on overall performance. The general trend shows that a larger λ_1 can enhance the consistency modelling between nodes at the same position, while a moderate λ_2 helps maintain the structural integrity of local regions. If λ_2 is too large, the relationships between local areas become overly smoothed, thereby weakening the discriminability of fine-grained features.

4.3.3 Comparison with state-of-the-art methods

Using the SYSU-MM01 and RegDB datasets, we do thorough comparisons with state-of-the-art techniques such as MUN, CAL, and DSAF in order to assess the efficacy of the suggested CSET model for cross-modal person re-identification. Table 3 provides a summary of the experimental findings.

Table 3 Comparison of the proposed CSET model with existing state-of-the-art methods (see online version for colours)

Methods	*SYSU-MM01*				*RegDB*			
	All-search		*Indoor-search*		*I2V*		*V2I*	
	R-1	*mAP*	*R-1*	*mAP*	*R-1*	*mAP*	*R-1*	*mAP*
FMCNet (Zhang et al., 2022)	66.34	62.51	68.15	74.09	88.38	83.86	89.12	84.43
SGIEL (Feng et al., 2023)	75.18	70.12	78.40	81.20	91.07	85.23	92.18	84.43
PMCM (Qian et al., 2025)	75.54	71.16	81.52	79.45	91.44	87.15	93.09	89.57
MUN (Yu et al., 2023)	76.24	73.81	79.42	82.06	91.86	85.01	95.19	87.15
MID (Huang et al., 2022)	60.27	59.40	64.86	70.12	84.29	81.41	87.45	84.85
CMTR (Liang et al., 2023)	65.45	62.90	71.46	76.67	84.92	80.79	88.11	81.66
PMT (Lu et al., 2023)	67.53	64.98	71.66	76.56	84.16	75.13	84.83	76.55
MSFCS (Yang et al., 2024)	70.59	67.49	75.98	80.24	83.88	75.16	85.34	76.39
MIP (Wu et al., 2024b)	70.84	66.41	78.80	79.92	92.38	85.99	91.26	85.90
CAJ+(Ye et al., 2023)	71.48	68.15	78.36	81.98	84.88	78.55	85.69	79.70
CAL(Wu et al., 2023)	74.66	71.73	79.69	83.68	93.64	87.61	94.51	88.67
DSAF(Jiang et al., 2025)	76.65	73.24	83.48	83.78	92.62	86.37	93.25	87.17
HTCR(Chen et al., 2025)	77.05	74.12	84.01	85.32	93.12	87.84	94.02	88.50
UIAL(Wei and Yin, 2025)	76.81	73.65	83.75	84.90	92.98	87.42	93.80	88.21
LEMF(Zhang et al., 2025)	77.26	74.58	84.39	85.61	93.24	87.95	94.15	88.73
CSET(Ours)	78.39	76.51	86.16	87.83	94.16	89.72	94.93	90.26

Note: Red indicates the best performance, and blue indicates the second-best.

From Table 2, it can be observed that under the all-search mode of the SYSU-MM01 dataset, CSET achieves Rank-1 accuracy and mAP of 78.39% and 76.51%, respectively, outperforming all compared methods. Under the indoor-search mode, CSET similarly achieves 86.16% Rank-1 accuracy and 87.83% mAP, demonstrating strong adaptability to complex indoor cross-modal matching scenarios.

On the RegDB dataset, CSET attains a Rank-1 accuracy of 94.16% and mAP of 89.72% in the I2V retrieval direction, achieving the best performance. In the V2I retrieval direction, CSET obtains the highest mAP of 90.26% among all methods, with a Rank-1 accuracy of 94.93%, which is slightly lower than the highest recorded 95.19%. Overall, CSET achieves breakthroughs in most evaluation metrics and strikes a better balance between retrieval accuracy and comprehensiveness.

Among existing methods, MUN alleviates modality discrepancy via auxiliary modality generation but struggles with precise local alignment due to its reliance on auxiliary features. CSET overcomes this by introducing a SE mechanism at each transformer layer, enhancing interaction between correlated cross-modal local features for better fine-grained alignment.

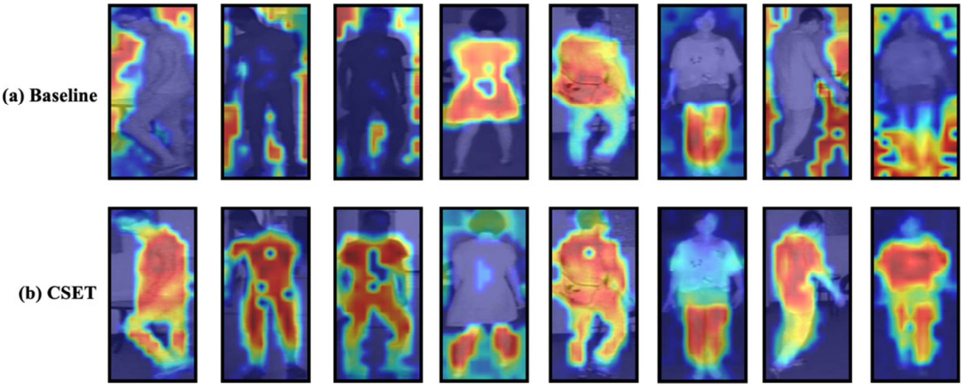
CAL improves semantic consistency across modalities through target-aware alignment but lacks structural modelling. CSET addresses this with a MHGA mechanism that captures both local and global relationships with adaptive importance weighting.

DSAF ensures identity consistency through dual-space alignment but models relational information coarsely. In contrast, CSET builds heterogeneous graphs at the deep feature level and adaptively adjusts adjacency based on positional embeddings, enabling more precise and robust cross-modal retrieval.

4.3.4 Visualisation experiments

To intuitively demonstrate CSET's advantages in cross-modal modelling and retrieval, we compare it with the transformer baseline by visualising feature response heatmaps and retrieval results, as shown in Figure 5.

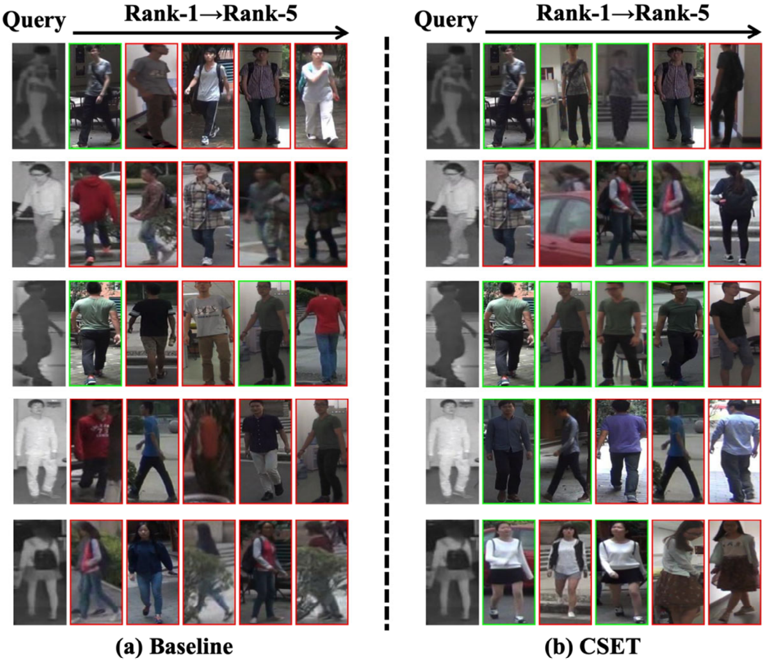
Figure 5 Isualisation comparison of feature response heatmaps extracted by CSET and the baseline (see online version for colours)



It can be observed that the feature response regions extracted by the baseline are relatively scattered, with a large number of high-response areas located in irrelevant background regions in the heatmaps. Moreover, attention to critical parts of the pedestrian body is insufficiently concentrated. Such fuzzy and dispersed feature representations lead to incomplete capture of local structural information, thereby reducing the discriminability and robustness of cross-modal matching. In contrast, the feature heatmaps extracted by CSET exhibit a more focused and coherent response pattern. CSET significantly highlights key parts of the pedestrian body, such as the head, upper body, and leg regions, while effectively suppressing background noise interference, demonstrating stronger local region perception capabilities.

Finally, we present the top-5 retrieval results based on the baseline and CSET to evaluate the models' accuracy and discriminative abilities in practical retrieval tasks. The comparison results are shown in Figure 6.

Figure 6 Visualisation comparison of top-5 retrieval results between CSET and the baseline (see online version for colours)



As observed in the top-5 ranked retrieval results, the baseline method produces many incorrect matches, particularly from Rank-1 to Rank-3, where the retrieved images often exhibit significant differences from the query target in posture, clothing, or background. This phenomenon indicates that the baseline model lacks sufficient discriminative ability in cross-modal retrieval tasks and is easily affected by modality differences and local feature deviations, leading to incorrect retrieval results. In contrast, the CSET method demonstrates much higher consistency and accuracy in the top-5 retrieval rankings. Most query images can retrieve the correct target identity within the top-3 ranks, and the other high-ranking retrieval results also maintain high appearance similarity.

5 Conclusions

To enhance the consistency and discriminability of modality features in cross-modal person re-identification (VI-ReID), we propose the CSET. By incorporating a SE mechanism and a MHGA mechanism, CSET improves fine-grained local feature alignment and cross-modal structural modelling. Experiments on the SYSU-MM01 dataset show CSET achieves Rank-1/mAP scores of 78.39%/76.51% (all-search) and 86.16%/87.83% (indoor-search), significantly surpassing state-of-the-art methods. On the RegDB dataset, it achieves Rank-1/mAP scores of 94.16%/89.72% (I2V) and 94.93%/90.26% (V2I), setting new benchmarks on most metrics. Ablation studies confirm the individual contributions of both proposed mechanisms, and sensitivity analysis demonstrates their robustness. Visualisation results further highlight CSET's strengths in feature localisation and retrieval accuracy.

Despite these improvements, some limitations remain. The current SE mechanism relies on static token similarity, which can become unstable under large pose variations, occlusions, or extreme lighting conditions. Similarly, the graph attention module, based on static positional embeddings, lacks adaptability to dynamic scenes. Future work will explore context-aware dynamic token matching and adaptive graph-based relationship modelling to further enhance the robustness and performance of the model in complex environments. Additionally, the impact of structural cue differences under partial occlusion or misalignment is another important direction for our future research. We will further investigate how to improve cross-modal feature alignment and structural modelling to address these complex scenarios.

Declarations

The datasets used and analysed during the current study available from the corresponding author on reasonable request.

The author declares that they have no conflicts of interest.

References

An, S., Chen, J., Xu, J., Kang, K. and Tang, R. (2024) 'Cross-modality transformer with mixed data augmentation learning for visible-infrared person re-identification', in *2024 9th International Conference on Cloud Computing and Big Data Analytics* (*ICCCBDA*), pp.168–175, IEEE.

Chang, H., Xu, X., Liu, W., Lu, L. and Li, W. (2024) 'A comprehensive survey of visible infrared person re-identification from an application perspective', *Multimedia Tools and Applications*, Vol. 83, No. 42, pp.90243–90270.

Chen, S., Lin, G., Hu, T., Wang, H. and Lai, Z. (2023a) 'Localization algorithm based on a spring particle model (LASPM) for large-scale unmanned aerial vehicle swarm (UAVs)', *International Journal of Cognitive Informatics and Natural Intelligence* (*IJCINI*), Vol. 17, No. 1, pp.1–13.

Chen, Z., Zhang, Z., Tan, X., Qu, Y. and Xie, Y. (2023b) 'Unveiling the power of clip in unsupervised visible-infrared person re-identification', in *Proceedings of the 31st ACM International Conference on Multimedia*, pp.3667–3675.

Chen, S., Qiu, L., Wang, D.H., Zhu, W., Hua, Y. and Yan, Y. (2025) 'Hierarchical token-aware cross-modality reconstruction for visible-infrared person re-identification', *IEEE Transactions on Multimedia*, pp.1–16.

Chong, J. (2023) 'An intelligent detection approach for smoking behavior', *International Journal of Cognitive Informatics and Natural Intelligence* (*IJCINI*), Vol. 17, No. 1, pp.1–18.

Fang, X., Yang, Y., and Fu, Y. (2023). Visible-infrared person re-identification via semantic alignment and affinity inference. In Proceedings of the IEEE/CVF international conference on computer vision. 11270-11279.

Feng, J., Wu, A. and Zheng, W. S. (2023) 'Shape-erased feature learning for visible-infrared person re-identification', in *Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition*, pp.22752–22761.

Feng, Y., Chen, F., Yu, J., Ji, Y., Wu, F., Liu, T., ... and Luo, J. (2024) 'Cross-modality spatial-temporal transformer for video-based visible-infrared person re-identification', *IEEE Transactions on Multimedia*, Vol. 26, pp.6582–6594.

Huang, B., Qi, X. and Chen, B. (2024) 'Cross-modal feature learning and alignment network for text–image person re-identification', *Journal of Visual Communication and Image Representation*, Vol. 103, p.104219.

Huang, N., Liu, J., Miao, Y., Zhang, Q. and Han, J. (2023) 'Deep learning for visible-infrared cross-modality person re-identification: a comprehensive review', *Information Fusion*, Vol. 91, pp.396–411.

Huang, Z., Liu, J., Li, L., Zheng, K. and Zha, Z. J. (2022) 'Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification', in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 1, pp.1034–1042.

Jiang, Y., Cheng, X., Yu, H., Liu, X., Chen, H. and Zhao, G. (2025) 'Dsaf: dual space alignment framework for visible-infrared person re-identification', *IEEE Transactions on Multimedia*, Vol. 27, pp.5591–5603.

Kim, M., Kim, S., Park, J., Park, S. and Sohn, K. (2023) 'Partmix: regularization strategy to learn part discovery for visible-infrared person re-identification', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.18621–18632.

Li, Y., Miao, D., Zhang, H., Zhou, J. and Zhao, C. (2024a) 'Multi-granularity cross transformer network for person re-identification', *Pattern Recognition*, Vol. 150, p.110362.

Li, Z., Wang, Q., Chen, L., Zhang, X. and Yin, Y. (2024b) 'Cascaded cross-modal alignment for visible-infrared person re-identification', *Knowledge-Based Systems*, Vol. 305, p.112585.

Liang, T., Jin, Y., Liu, W. and Li, Y. (2023) 'Cross-modality transformer with modality mining for visible-infrared person re-identification', *IEEE Transactions on Multimedia*, Vol. 25, pp.8432–8444.

Liu, J., Bai, W. and Hui, Y. (2024) 'Reverse pyramid attention guidance network for person re-identification', *International Journal of Cognitive Informatics and Natural Intelligence* (*IJCINI*), Vol. 18, No. 1, pp.1–22.

Lu, H., Zou, X. and Zhang, P. (2023) 'Learning progressive modality-shared transformers for effective visible-infrared person re-identification', in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 2, pp.1835–1843.

Nguyen, D.T., Hong, H.G., Kim, K.W. and Park, K.R. (2017) 'Person recognition system based on a combination of body images from visible light and thermal cameras', *Sensors*, Vol. 17, No. 3, p.605.

Pan, H., Pei, W., Li, X. and He, Z. (2024) 'Unified conditional image generation for visible-infrared person re-identification', *IEEE Transactions on Information Forensics and Security*, Vol. 19, pp.9026–9038.

Pang, Z., Wang, C., Zhao, L., Liu, Y. and Sharma, G. (2023) 'Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 34, No. 4, pp.2706–2718.

Qian, Z., Lin, Y. and Du, B. (2025) 'Visible–infrared person re-identification via patch-mixed cross-modality learning', *Pattern Recognition*, Vol. 157, p.110873.

Sarker, P.K. and Zhao, Q. (2024) 'Enhanced visible–infrared person re-identification based on cross-attention multiscale residual vision transformer', *Pattern Recognition*, Vol. 149, p.110288.

Sarker, P.K., Zhao, Q. and Uddin, M.K. (2024) 'Transformer-based person re-identification: a comprehensive review', *IEEE Transactions on Intelligent Vehicles*, Vol. 9, No. 7, pp.5222–5239.

Shi, J., Yin, X., Chen, Y., Zhang, Y., Zhang, Z., Xie, Y. and Qu, Y. (2024) 'Multi-memory matching for unsupervised visible-infrared person re-identification', in *European Conference on Computer Vision*, pp.456–474, Springer Nature Switzerland, Cham.

Shi, J., Zhang, Y., Yin, X., Xie, Y., Zhang, Z., Fan, J., ... and Qu, Y. (2023) 'Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.11218–11228.

Wei, C. and Yin, B. (2025) 'Unified identity and attribute learning for visible-infrared person re-identification', in *International Conference on Intelligent Computing*, July, pp.255–266, Springer Nature Singapore, Singapore.

Wei, Z., Yang, X., Wang, N. and Gao, X. (2023) 'Dual-adversarial representation disentanglement for visible infrared person re-identification', *IEEE Transactions on Information Forensics and Security*, Vol. 19, pp.2186–2200.

Wu, A., Zheng, W.S., Yu, H.X., Gong, S. and Lai, J. (2017) 'RGB-infrared cross-modality person re-identification', in *Proceedings of the IEEE International Conference on Computer Vision*, pp.5380–5389.

Wu, J., Liu, H., Su, Y., Shi, W. and Tang, H. (2023) 'Learning concordant attention via target-aware alignment for visible-infrared person re-identification', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.11122–11131.

Wu, R., Jiao, B., Wang, W., Liu, M. and Wang, P. (2024a) 'Enhancing visible-infrared person re-identification with modality-and instance-aware visual prompt learning', in *Proceedings of the 2024 International Conference on Multimedia Retrieval*, May, pp.579–588.

Wu, T., Zhang, S., Chen, D. and Hu, H. (2024b) 'Text-and-image learning transformer for cross-modal person re-identification', *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 21, No. 1, pp.1–18.

Wu, Z. and Ye, M. (2023) 'Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9548–9558.

Yang, X., Dong, W., Li, M., Wei, Z., Wang, N. and Gao, X. (2024) 'Cooperative separation of modality shared-specific features for visible-infrared person re-identification', *IEEE Transactions on Multimedia*, Vol. 26, pp.8172–8183.

Yang, X., Wang, J., Sun, Y. and Duan, X. (2025) 'CMLFA: cross-modal latent feature aligning for text-to-image person re-identification', *Journal of Electronic Imaging*, Vol. 34, No. 1, p.13018.

Ye, M., Wu, Z., Chen, C. and Du, B. (2023) 'Channel augmentation for visible-infrared re-identification', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 4, pp.2299–2315.

Yu, H., Cheng, X., Peng, W., Liu, W. and Zhao, G. (2023) 'Modality unifying network for visible-infrared person re-identification', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.11185–11195.

Zhai, Y., Zeng, Y., Huang, Z., Qin, Z., Jin, X. and Cao, D. (2024) 'Multi-prompts learning with cross-modal alignment for attribute-based person re-identification', in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 7, pp.6979–6987.

Zhang, L., Zhao, X., Du, H., Sun, J. and Wang, J. (2025) 'Learning enhancing modality-invariant features for visible-infrared person re-identification', *International Journal of Machine Learning and Cybernetics*, Vol. 16, No. 1, pp.55–73.

Zhang, Q., Lai, C., Liu, J., Huang, N. and Han, J. (2022) 'Fmcnet: feature-level modality compensation for visible-infrared person re-identification', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7349–7358.

Zhang, Y. and Wang, H. (2023) 'Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.2153–2162.

Zhang, Y., Yan, Y., Li, J. and Wang, H. (2023) 'MRCN: a novel modality restitution and compensation network for visible-infrared person re-identification', in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 3, pp.3498–3506.

Zheng, X., Huang, X., Ji, C., Yang, X., Sha, P. and Cheng, L. (2024) 'Multi-modal person re-identification based on transformer relational regularization', *Information Fusion*, Vol. 103, p.102128.