Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062895

ARTICLE

Visible-Infrared Person Re-Identification via Quadratic Graph Matching and Block Reasoning

Junfeng Lin[1], Jialin Ma[1,*], Wei Chen[1,2], Hao Wang[1], Weiguo Ding[1] and Mingyao Tang[1]

[1]Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian, 223003, China
[2]Jiangsu Suyan Jingshen Co., Ltd., Huaian, 223003, China

*Corresponding Author: Jialin Ma. Email: majl@hyit.edu.cn

Received: 30 December 2024; Accepted: 26 March 2025; Published: 09 June 2025

ABSTRACT: The cross-modal person re-identification task aims to match visible and infrared images of the same individual. The main challenges in this field arise from significant modality differences between individuals and the lack of high-quality cross-modal correspondence methods. Existing approaches often attempt to establish modality correspondence by extracting shared features across different modalities. However, these methods tend to focus on local information extraction and fail to fully leverage the global identity information in the cross-modal features, resulting in limited correspondence accuracy and suboptimal matching performance. To address this issue, we propose a quadratic graph matching method designed to overcome the challenges posed by modality differences through precise cross-modal relationship alignment. This method transforms the cross-modal correspondence problem into a graph matching task and minimizes the matching cost using a center search mechanism. Building on this approach, we further design a block reasoning module to uncover latent relationships between person identities and optimize the modality correspondence results. The block strategy not only improves the efficiency of updating gallery images but also enhances matching accuracy while reducing computational load. Experimental results demonstrate that our proposed method outperforms the state-of-the-art methods on the SYSU-MM01, RegDB, and RGBNT201 datasets, achieving excellent matching accuracy and robustness, thereby validating its effectiveness in cross-modal person re-identification.

KEYWORDS: Cross-modal; person re-identification; modal correspondence; quadratic graph matching; block reasoning

1 Introduction

Person re-identification (ReID) is the task of retrieving pedestrian images captured by different cameras. Existing ReID methods primarily focus on matching visible light images [1,2], addressing the unimodal problem. However, conventional visible surveillance cameras cannot effectively capture pedestrian information under poor lighting conditions [3,4]. To address this challenge, modern surveillance cameras can automatically switch to infrared mode to capture infrared images under low-light conditions [5]. Consequently, the research on visible-infrared (VI) person re-identification has emerged, with the goal of identifying the same person from a visible/infrared image database when given an image from the other modality. Due to the importance of this task in nighttime intelligent monitoring and public safety, it has recently gained widespread attention [6,7], and significant progress has been made in the VI-ReID field.

However, most existing methods [8,9] perform recognition by mapping features to a shared space, which often results in poor performance due to feature loss—critical modality-specific attributes are suppressed

during alignment, leading to ambiguous representations. Therefore, we aim to explore a reliable cross-modal person re-identification solution that preserves modality-invariant identity cues without sacrificing discriminative details.

Recently, some studies have focused on finding correspondences between different modalities [10,11]. However, most methods, as shown in Fig. 1, tend to extract common features for modality alignment, which often leads to the loss of local information and fails to fully utilize the global information between different identities. Additionally, regarding the issue of cluster imbalance, many methods [12,13] discard certain clusters when correspondences cannot be found, further increasing the gap between modalities. To address this, we propose a quadratic graph matching (QGM) method to prevent local information loss and make full use of clustering results. This method primarily connects the two modalities through graph matching and adopts a quadratic matching strategy to tackle the cluster imbalance problem.

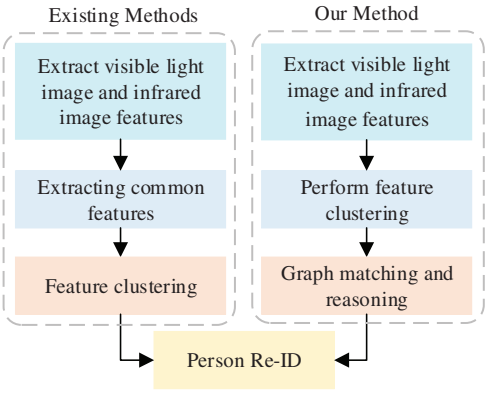

Figure 1: The difference between our method and existing methods

First, we fully leverage the relationships between different identities using graph matching, which is processed under global constraints. This approach transforms the process of discovering cross-modal correspondences into a bipartite graph matching problem, where each modality is viewed as a graph, and each cluster's representative sample is treated as a node. The matching cost between nodes is positively correlated with the distance between clusters. By minimizing the global matching cost, graph matching generates more reliable global correspondences rather than local feature alignments. A large body of research [14–16] has demonstrated the advantages of graph matching in establishing correspondences between feature sets. Inspired by this, we construct graphs for each modality and connect the same person across different modalities.

Basic graph matching struggles to solve the cross-modal cluster imbalance problem. To address this, we propose a quadratic matching strategy. Due to variations in camera settings, similar samples may be assigned to different clusters, and these new clusters lack correspondences, which affects the reduction of modality gaps. By using dynamic quadratic matching, we progressively find correspondences for each cluster. Subgraphs of the bipartite graph in one matching process are continuously updated based on previous matching results until each cluster finds a correspondence. Through this strategy, the same identity in different clusters can find the same cross-modal correspondence, thereby solving the imbalance issue while enhancing intra-class compactness.

To speed up pedestrian image retrieval and updates, we also propose a Block Reasoning (BR) module, which can more efficiently utilize the affinity information between images. Specifically, we first partition the

database images according to bipartite graph nodes and combine these node images to form new database images. Furthermore, similar to most existing methods [17–19], we compute affinity matrices for query-database and database-database pairs, and use these two matrices to adjust the measured distance. By finding the matching nodes based on the distance, we can then identify the pedestrian's identity based on the node's class. At the same time, the block structure allows for efficient updating of the database images through image comparisons.

Our main contributions are summarized as follows:

- We propose the QGM method for mining reliable cross-modal correspondences in VI-ReID. First, modality graphs are constructed and graph matching is performed to integrate global information between identities. Then, a quadratic matching strategy is applied to address the cluster imbalance problem, making the matching process more adaptive.
- We introduce the BR module, which not only enables efficient matching using the relationships between pedestrians but also facilitates dynamic data updates.
- Comprehensive experimental results validate the effectiveness of the proposed framework. Under various test conditions, the performance of the proposed method outperforms the state-of-the-art methods.

2 Related Work

2.1 Visible-Infrared Person ReID

The visible-infrared (VI) person re-identification task focuses on matching pedestrian images captured by visible light and infrared cameras. Existing methods can mainly be categorized into two types based on feature processing approaches: generative methods and non-generative methods. Generative Methods: These methods [20] focus on reducing the style differences between modalities. The mainstream approach is to use Generative Adversarial Networks (GANs) for modality translation. For example, MUNIT-GAN [15] and AttGAN [21] utilize GANs to perform unsupervised image-to-image translation across multiple modalities. However, these methods often increase the computational load of the model and may introduce additional noise. In contrast, non-generative methods directly exploit raw features without data synthesis. Recent advances include semantic-driven frameworks like CLIP-Driven [22], which align cross-modal semantics using vision-language models but require costly text annotations; spatio-temporal aggregation techniques [23] that enhance temporal consistency in video sequences at the expense of high computational complexity; and multi-view clustering approaches [24,25], which complete incomplete multi-view data via tensor decomposition or manifold learning but rely on the restrictive assumption of aligned inputs. To address these limitations, the proposed method achieves unsupervised cross-modal alignment through quadratic graph matching and block reasoning, eliminating the need for manual annotations or complex fusion frameworks.

2.2 Graph Matching for Person ReID

In unimodal ReID, graph matching is mainly used in two ways: one is to divide pedestrian images into multiple parts, treating local features as nodes in the graph and using graph matching to align features of pedestrians under different poses and occlusion scenarios [26,27]. The other is to model pedestrian images as a graph structure, where nodes represent feature points and edges represent the relationships between features. Graph matching is used to analyze and compare the structural similarity between different images [28,29]. However, in VI-ReID, the cross-modal differences are much larger than the cross-camera differences within a single modality. Therefore, we construct a graph for each modality and use graph matching to discover the cross-modal correspondences. Recent work [30] proposed progressive graph

matching for VI-ReID. However, their method relies on alternating optimization between feature learning and graph matching, which may lead to suboptimal convergence. In contrast, our quadratic graph matching dynamically resolves cluster imbalance through iterative subgraph updates, ensuring that all clusters find correspondences without discarding outliers. Additionally, we integrate a center search mechanism to select representative nodes with minimal variance, further enhancing matching stability.

2.3 Inference Methods for VI-ReID

Most existing methods [31,32] perform inference using simple distance metrics, such as calculating Euclidean or Mahalanobis distances between output features to measure similarity. Although these methods are simple and intuitive, they treat each database image as an independent entity and ignore the potential relationships between database images, leading to the loss of valuable affinity data, which in turn affects the matching performance. To address this issue, the similarity inference metric proposed in [33] incorporates the calculation of Jaccard distance between database images to optimize the matching. However, Jaccard distance only considers the presence or absence of elements and ignores specific similarity scores, thus limiting its effectiveness to some extent. In response, the affinity inference metric proposed in [34] considers the similarity between database images but requires recalculating large amounts of data when querying or updating the database images. To address this, we propose the BR module.

3 Methodology

This chapter provides a detailed description of the method we propose. The overall framework of the method is shown in Fig. 2). We utilize the Dual Contrastive Learning (DCL) framework (on the left side of Fig. 2) to learn discriminative features within each modality and optimize them using modality-specific contrastive loss functions. Furthermore, based on DCL, we introduce the innovative methods presented in this paper, focusing on the novel QGM module (in the center of Fig. 2) and the BR module during the testing phase (on the right side of Fig. 2). The QGM module consists of two parts: the center search, which selects representative points that are not affected by outliers by analyzing the relationships between sample points and cluster centers using variance; and the construction of a graph to establish modality correspondence. The BR module leverages the matching results from graph matching to partition the image gallery, calculating the affinity distance between blocks to achieve more accurate person re-identification. Additionally, it uses the advantages of blocks to enable fast updating of the image gallery. The detailed descriptions of these two modules are provided in Sections 3.2 and 3.3.

3.1 Dual-Contrastive Learning Framework

To clearly describe the method, let $T = \{T^v, T^r\}$ represent the training dataset of visible-infrared images, where $T^v = \{x_i^v | i = 1, 2, \cdots, N\}$ refers to the visible dataset consisting of N visible images, and $T^r = \left\{x_j^r \middle| j = 1, 2, \cdots, M\right\}$ refers to the infrared dataset consisting of M infrared images. Grayscale augmentation is a common data augmentation technique for visible light images, aiming to remove color information and retain only brightness (intensity) information. Therefore, grayscale augmentation is suitable for the learning process of visible light images. For infrared image learning, temperature mapping augmentation is more appropriate, as it expands the infrared representation based on temperature values from different regions.

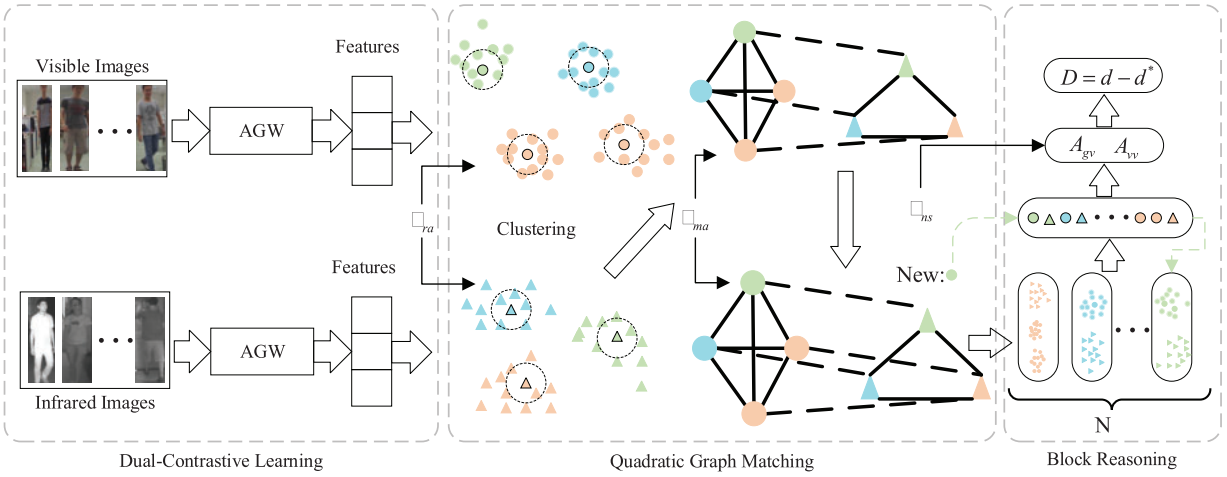

Figure 2: The pipeline of our framework. Different colors indicate different pedestrians

We use AGW as our feature extraction framework, which is based on the ResNet50 backbone to extract features from visible and infrared images, respectively. The extracted feature sets are then clustered using DBSCAN. Subsequently, the contrastive loss functions in Eq. (1) are applied to train the model separately on visible and infrared clusters.

$$\mathcal{L}_{ra} = y_{ij} \cdot D\left(F_v^i, F_r^j\right) + \left(1 - y_{ij}\right) \cdot \max\left(0, m - D\left(F_v^i, F_r^j\right)\right). \tag{1}$$

Let $\mathcal{L}_{ra}$ represent the unimodal loss function; F_v^i denotes the feature vector corresponding to sample x_i^v, and F_r^j denotes the feature vector corresponding to sample x_j^r. $D\left(F_v^i, F_r^j\right)$ represents the Euclidean distance between the two features. y_{ij} is a binary label indicating whether the sample pair belongs to the same category (1 for the same category, 0 for different categories). The loss function consists of two parts: when $y_{ij} = 1$, we directly minimize the Euclidean distance; when $y_{ij} = 0$, we enforce a margin constraint to ensure the distance is at least m. After completing the training, we achieve the unsupervised person re-identification clustering task within each modality.

3.2 Quadratic Graph Matching

In the DCL module described above, we did not directly address the relationship between the two modalities. Therefore, when the gap between the two modalities is significant, the above method cannot be applied. To address the cross-modal correspondence problem, we propose the QGM module. Prior to matching, we first perform center search, which serves as the foundation for graph matching. The feature distribution results after quadratic graph matching are shown in Fig. 3.

The Center Search Module is designed to select the most representative sample points from each cluster, overcoming the interference of outliers and enhancing the robustness of cross-modal alignment. Its core idea is to choose the point that is most consistent with the sample distribution of the same class, based on the stability (variance) of the distance distribution rather than just geometric distance. Compared to traditional methods, such as using K-means to directly select centroids, this approach better adapts to noisy data distributions.

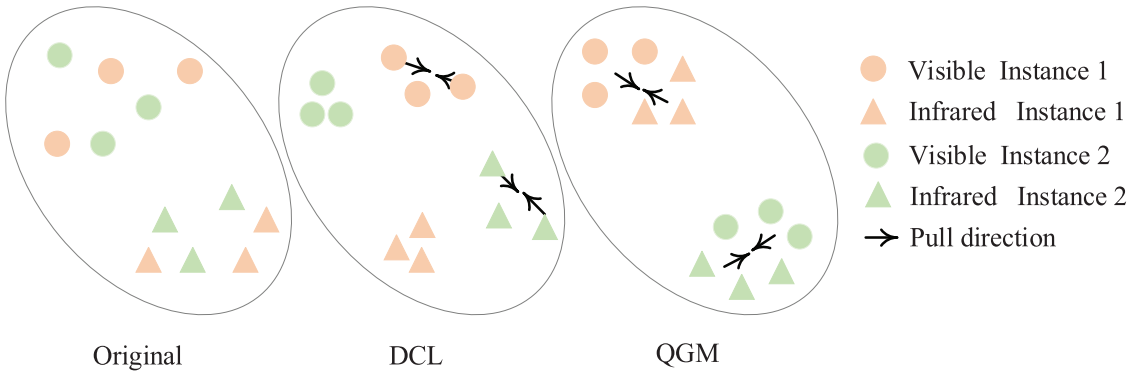

Figure 3: The effect of the original feature after DCL and QGM processing

First, the initial distances are calculated, and candidate points are selected. For a given cluster, we compute the Euclidean distance from each point to the current cluster centroid and sort these distances in ascending order. The top ten sample points with the smallest distances form the candidate point set $X = \{x_k\}_{k=1}^{L}$ for the next round of filtering. If the number of points in the cluster is fewer than ten but greater than two, the actual number of sample points is used for the next round of filtering. If the number of points in the cluster is less than two, the variance calculation step is skipped, and only the point closest to the centroid is selected. Next, the distance distribution between the candidate points is calculated. For each candidate point, we compute its Euclidean distance to all other candidate points within the cluster and then calculate the mean and variance of these distances. The specific expressions are shown in Eqs. (2) and (3).

$$V_k = \frac{1}{N-1} \sum_{k \neq l} \left(D\left(x_k, x_l\right) - MD_k \right)^2, \tag{2}$$

$$MD_k = \frac{1}{N-1} \sum_{k \neq l} D\left(x_k, x_l\right), \tag{3}$$

where V_k denotes the variance of point x_k. N represents the actual number of selected sample points. $D\left(x_k, x_l\right)$ represents the Euclidean distance between point x_k and point x_l. MD_k denotes the average distance from point x_k to all other selected points. After calculating the variance for each point, we select the point with the smallest variance as the representative point for the cluster. This is because the point with the smallest variance best represents the position of all points in the cluster. It has the most stable distance to other points and is less affected by outliers.

Furthermore, we construct two graphs, $G_{vis} = (V_{vis}, E_{vis})$ and $G_{ir} = (V_{ir}, E_{ir})$, where the node sets represent the selected representative visible and infrared image features, denoted as $V_{vis} = \left\{ f_v^i \middle| i = 1, 2, \cdots, K \right\}$ and $V_{ir} = \left\{ f_r^j \middle| i = 1, 2, \cdots, L \right\}$, respectively. The edge sets, E_{vis} and E_{ir}, represent the similarity between features. We compute the similarity matrix, S, between visible and infrared features, where the element at the i row and j column of the matrix is denoted as $s_{ij} = sim\left(f_v^i, f_r^j\right)$, calculated using cosine similarity. Based on the similarity matrix, S, we traverse each row and column and select feature pairs with a similarity greater than 0.75 as the initial matching pairs. Based on the initial matching results, we construct new graphs, G_{vis}^{local} and G_{ir}^{local}, by combining the remaining visible and infrared image features. The process of graph construction follows the same steps as the first construction. We then recalculate the similarity between visible and infrared features to obtain a new similarity matrix, S', where the element at the i row and j column of the new matrix can be represented by Eq. (4).

$$s'_{ij} = \alpha \cdot sim\left(f_v^i, f_r^j\right) + \beta \cdot struct\_sim\left(G_{vis}^{local}, G_{ir}^{local}\right), \tag{4}$$

where α and β are weight parameters, and $struct\_sim$ represents the structural similarity, specifically using the Structural Similarity Index. Based on the updated similarity matrix, S', we compare the number of rows and columns. Assuming there are fewer rows, we traverse each row and select the most similar feature in each row as a matching pair, until each row has a corresponding match. The same procedure is followed if there are fewer columns.

After this node matching process, we assume that there are remaining nodes in graph G_{vis} that have not found corresponding nodes, while all nodes in graph G_{ir} have corresponding matches. We dynamically rebuild a new graph, G'_{vis}, using the remaining nodes in G_{vis} and the edges between them. The new edge set G'_{vis} and G_{ir} are reorganized into a bipartite graph. We then reapply the node matching procedure on the reorganized bipartite graph. At this point, there is no need to compare the number of rows and columns. If there are remaining nodes in graph G_{vis} without a corresponding match, we only need to find matching pairs for the remaining nodes in graph G'_{vis} to conclude the task. Finally, ensure that every node in both sets G_{vis} and G_{ir} finds its corresponding match. During the entire matching process, the goal is not only to ensure that every class finds its corresponding class but also to minimize incorrect matches and enhance the accuracy of matching. Based on this approach, we further propose a modality association loss function, $\mathcal{L}_{ma}$, which can be expressed as Eq. (5).

$$\mathcal{L}_{ma} = \sum_{i}^{K} \sum_{j}^{L} \max\left(0, \gamma - sim\left(f_v^i, f_r^j\right)\right), \tag{5}$$

where γ represents a threshold for controlling the minimum similarity requirement, and $sim\left(f_v^i, f_r^j\right)$ represents the cosine similarity between visible landmark points i and infrared landmark points j.

3.3 Block Reasoning

During the testing phase, pedestrian image matching based on Euclidean or cosine distance is quite limited. While Euclidean distance directly measures the geometric distance between features and cosine distance evaluates angular similarity, both approaches assume that feature relationships are independent, ignoring contextual correlations among samples. This limitation results in reduced robustness under modality variations. These methods often overlook the relationships between images or treat the entire image gallery as a whole, leading to excessive computational overhead during image updates. To overcome these limitations, we introduce the BR module, which dynamically constructs affinity groups based on the relational structure of samples rather than relying solely on pairwise distance metrics. Specifically, we can adjust the distance by utilizing gallery images that exhibit high affinity with the query image. BR captures the latent affinity information between gallery images and incorporates it into the distance calculation, thereby optimizing the matching performance. The detailed distance updating process is shown in Fig. 4.

Based on the graph matching results mentioned earlier, we group the matched nodes corresponding to the visible and infrared datasets into blocks, ultimately forming N blocks. When the number of visible nodes exceeds the number of infrared nodes, the number of blocks, N, represents the number of infrared nodes. Conversely, if the number of visible nodes is fewer than the infrared nodes, N represents the number of visible nodes. In addition, we construct two affinity matrices: one is the affinity matrix, A_{gv}, consisting of the matched node image sets from graph v and the query gallery g, and the other is the affinity matrix, A_{vv}, constructed from the matched node image sets from graph v.

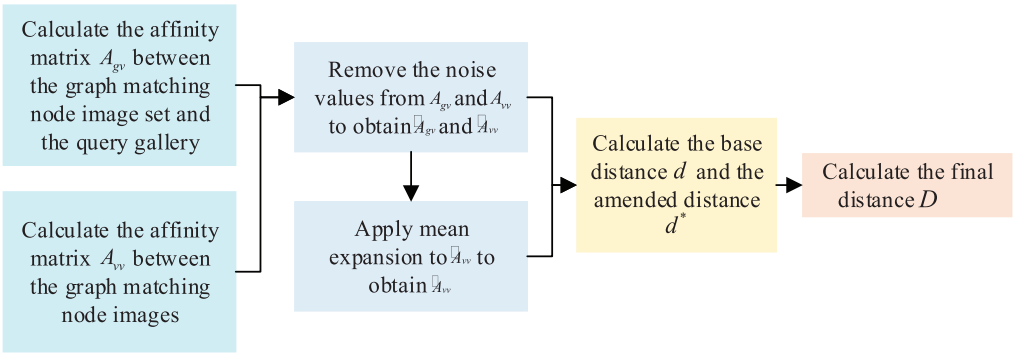

Figure 4: Distance update process diagram

We compute the affinity matrix, A_{gv}, between the graph-matching node image sets and the query gallery, where each element, $A_{gv}(i,j)$, represents the cosine similarity between node image v_i and query image g_j, the specific expression is shown in Eq. (6).

$$A_{gv}(i,j) = 1 - \frac{F_{v_i} \cdot F_{g_j}}{\|F_{v_i}\| \|F_{g_j}\|}, \tag{6}$$

where F_{v_i} represents the feature of the node image v_i, and F_{g_j} represents the feature of the query image g_j. Next, we compute the affinity matrix, A_{vv}, between the graph-matching node images, where each element, $A_{vv}(j,k)$, represents the similarity between image v_j and image v_k, the specific expression is shown in Eq. (7).

$$A_{vv}(j,k) = \begin{cases} 1 - \dfrac{F_{v_j} \cdot F_{v_k}}{\|F_{v_j}\| \|F_{v_k}\|} & j \neq k \\ 1 & j = k \end{cases}. \tag{7}$$

The dual-contrastive loss aims to bring together images of the same identity while pushing away those with low similarity. However, the associative information between low-similarity images is often too weak to be effectively utilized and may introduce noise, which interferes with subsequent distance calculations, leading to inaccurate results. Therefore, we need to eliminate these noise values.

To address this issue, we introduce a noise suppression mechanism to reduce the impact of inaccurate affinity values on matching. We define a threshold, θ, and set affinity values below this threshold to zero. This operation are applied to affinity matrix A_{gv} and A_{vv}. The affinity matrix after noise suppression is represented as $\tilde{A}_{gv}$ and $\tilde{A}_{vv}$, and for the matrix $\tilde{A}_{gv}$, the specific noise suppression expression is shown in Eq. (8).

$$\tilde{A}_{gv}(i,j) = \begin{cases} A_{gv}(i,j) & A_{gv}(i,j) \geq \theta \\ 0 & A_{gv}(i,j) < \theta \end{cases}. \tag{8}$$

To further improve the suppression of inaccurate similarity information during the task execution and ensure the model focuses on meaningful associations, we propose a noise suppression loss function, $\mathcal{L}_{ns}$, the function can be represented by Eq. (9).

$$\mathcal{L}_{ns} = \sum_{i}^{K} \sum_{j}^{L} \mathbb{I}_{(p_{ij} < \varepsilon)} \cdot \left(p_{ij}\right)^2, \tag{9}$$

where p_{ij} represents the affinity value, i.e., the element at row i and column j of the affinity matrix, and $\mathbb{I}_{(p_{ij}<\varepsilon)}$ is the indicator function, which is 1 when the condition is satisfied and 0 otherwise. Further, mean expansion is required to reduce the distance between images. Specifically, for a given gallery image, we can find M most similar images, and then replace the affinity values in the image with the average affinity of these M images. This operation applies only to the affinity matrix $\tilde{A}_{vv}$, and the specific mean expansion can be expressed as Eq. (10).

$$\hat{A}_{vv}(i,j) = \begin{cases} \tilde{A}_{vv}(i,j) & \tilde{A}_{vv}(i,j) = 0 \\ A_{mean\_row}(i) & \tilde{A}_{vv}(i,j) \neq 0 \end{cases}, \tag{10}$$

where $A_{mean\_row}(i)$ represents the average affinity from the M images, and $A_{mean\_row}(i) = \frac{1}{M}\sum_{j=1}^{M}\tilde{A}_{vv}(i,j)$ is the expression for the average affinity. The affinity matrix $\hat{A}_{vv}$ represents the matrix $\tilde{A}_{vv}$ after mean expansion. The expansion process does not require checking whether the element is zero; instead, we find the M most similar images and replace the affinity value with the average affinity of these images.

Initially, we use cosine similarity for A_{gv}, which can be transformed into base distance, as represented by Eq. (11).

$$d = 1 - A_{gv}. \tag{11}$$

The affinity reasoning module works such that if a query image is similar to a node image, the distance between the query image and the similar node images should be reduced. The distance reduction depends on the distances between the query image and the node image's similar images. Therefore, the corrected distance between each query image and each node image can be represented by Eq. (12).

$$d^* = \tilde{A}_{gv}\hat{A}_{vv}. \tag{12}$$

Finally, by subtracting the corrected distance between images from the base distance, D, the final affinity distance between images can be obtained, which is represented by Eq. (13).

$$D = d - d^*. \tag{13}$$

Based on the final distance, we query the two closest images for any given query image. If both images belong to the same data block, we classify the query image under that block's label. If the two images come from different data blocks, we further compute the average cosine similarity between the query image and each of the two blocks, and the block with the higher score determines the label identity for the query image. For data updates, new data only needs to be compared with the node image dataset, and classification is done based on the scoring results.

3.4 Training and Inference

The overall loss function $\mathcal{L}$ of QGM-BR can be expressed as Eq. (14).

$$\mathcal{L} = \mathcal{L}_{ra} + \mathcal{L}_{ma} + \lambda\mathcal{L}_{ns}, \tag{14}$$

where λ is a hyperparameter used to balance the contribution of the loss term $\mathcal{L}_{ns}$.

4 Experiments

In this section, we first introduce the datasets and experimental details. Then, experiments are conducted on two publicly available datasets. Finally, a detailed analysis of QGM-BR is presented.

4.1 Datasets

We evaluate the proposed method on three widely used visible-infrared datasets, SYSU-MM01 [35], RegDB [36] and RGBNT201 [37].

SYSU-MM01 Dataset. This dataset contains images of 1900 pedestrians from six different viewpoints. Each pedestrian has images captured from two visible light viewpoints, along with one thermal infrared viewpoint. It is one of the most challenging datasets for cross-modal person re-identification. Testing on the SYSU-MM01 dataset is conducted under two settings: full retrieval mode and indoor retrieval mode. In full retrieval mode, the gallery consists of visible light images, while the query set consists of infrared images. In indoor retrieval mode, visible light images from outdoor scenes (cameras 4 and 5) are excluded.

RegDB Dataset. This dataset contains 412 identities, each with 10 RGB images and 10 thermal images. The dataset includes 254 females and 158 males. Among the 412 identities, 156 were captured from the front, and 256 from the back. The RegDB dataset contains two testing settings: infrared to visible light and visible light to infrared modes.

RGBNT201 Dataset. This dataset is a pedestrian image database that includes three modalities: visible light, infrared, and thermal imaging. According to the original data split, the training subset consists of 141 classes (3280 visible light images and 3280 infrared images), while the test subset consists of 30 classes (836 visible light images and 836 infrared images). In practice, we only use the visible light and infrared images from each class for experimentation. Similar to the evaluation on the RegDB dataset, two retrieval modes are used: Visible to Thermal and Thermal to Visible. In the Visible to Thermal retrieval mode, the probe set is constructed by randomly selecting 10 visible light images from each class in the test set, while the gallery set contains all infrared images from the test set. The Thermal to Visible retrieval mode has a similar probe and gallery structure, but with the modality configuration reversed. For both retrieval modes, the final results are reported as the average of ten tests.

4.2 Evaluation Protocols

4.2.1 Implementation Details

During the training phase, we use a non-local module enhanced network based on AGW [14], with ResNet50 [38] as the feature extractor. The backbone network is initialized with pre-trained weights from ImageNet. In each mini-batch, the number of classes P and the number of samples per class K are both set to 8. All pedestrian images are resized to 256 × 128 pixels. The model is trained using the SGD optimizer, with an initial learning rate of 0.1 for randomly initialized parameters and 0.01 for pre-trained parameters. The learning rate is reduced by a factor of 10 at epochs 20 and 50. Random grayscale augmentation is applied to visible light images. During each training epoch, DBSCAN is used to cluster images within each modality. The maximum distance is set to 0.5 for the SYSU-MM01 dataset and 0.25 for the RegDB dataset. The minimum cluster size for both datasets is set to 4. During clustering, the memory update rate λ is set to 0.05, the temperature factor τ is set to 0.1, and the weight parameter μ is set to 0.5. All experiments are conducted on an NVIDIA RTX 4090D GPU.

4.2.2 Evaluation Metrics

During the experimental phase, we use standard evaluation metrics to assess the two datasets, including the Cumulative Match Characteristic (CMC) curve, Mean Inverse Negative Penalty (mINP), and Mean Average Precision (mAP) to measure the model's recognition performance. The CMC curve reflects the classification accuracy of the model, typically represented as Rank-n for the top n matching results. mINP indicates the proportion of correct samples among all retrieved samples up to the last correct match. mAP is the mean accuracy of all returned results for a given category.

4.3 Parameters Analysis

To find the optimal hyperparameters for the proposed method, we first conducted a parameter analysis experiment to examine the impact of the weighted combination of different loss functions on the model. As shown in Fig. 4. Our study aims to verify the impact of the proposed BR module on overall performance. This experiment was conducted in the global mode of the SYSU-MM01 dataset, testing the model across the parameter range $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. When the parameter is 0, the BR module is not used, resulting in poor model performance. The final experimental results, shown in Fig. 5, indicate that the proposed BR module effectively improves the model's performance, with optimal results achieved at a specific parameter value.

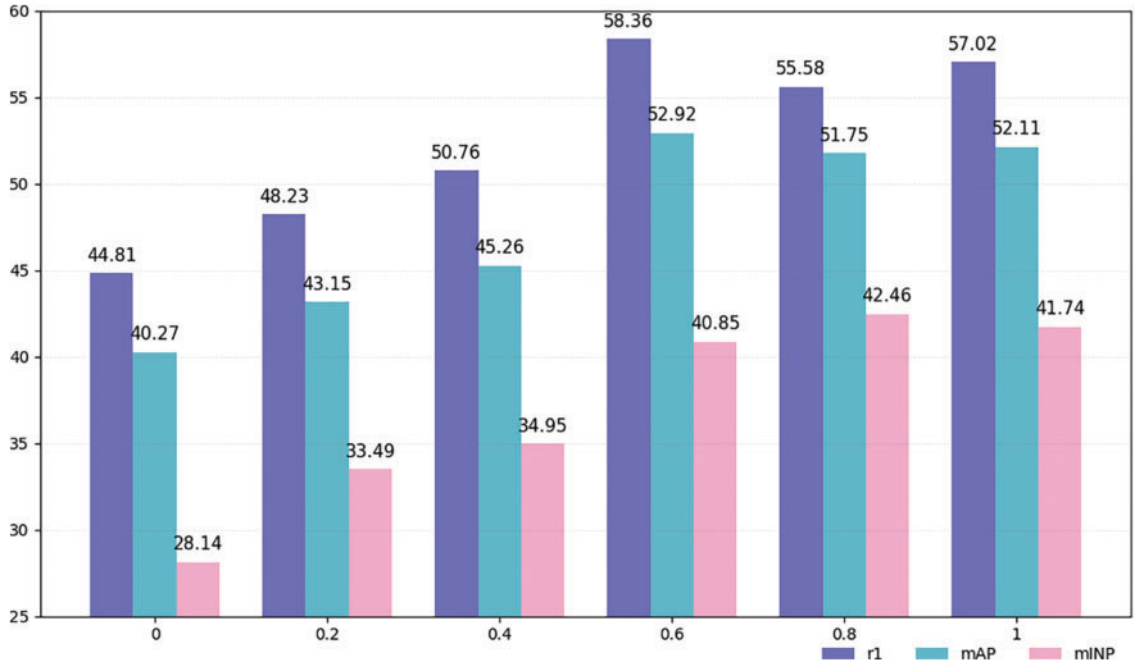

Figure 5: The impact of different λ values on the SYSU-MM01 dataset under the all-search mode

4.4 Comparison with State-of-the-Art Methods

Based on the optimal model derived in the previous section, we first evaluate the framework we proposed using the widely used SYSU-MM01 and RegDB datasets. The comparison methods are mainly divided into two categories: one for unsupervised cross-modal visible-infrared person re-identification methods; and another for unsupervised single-modality person re-identification methods. The comparison results are shown in Table 1.

Table 1: Comparison of experimental results of different methods on the SYSU-MM01 dataset (%)

Methods	Reference	All search				Indoor search			
		Rank-1	Rank-10	mAP	mINP	Rank-1	Rank-10	mAP	mINP
SPCL [39]	NIPS 2020	18.37	54.08	19.39	10.99	26.83	68.31	36.42	33.05
AGW [14]	TPAMI 2021	47.5	84.39	47.65	35.3	54.17	91.94	62.97	59.23
MMT [40]	ICLR 2020	21.47	59.65	21.53	11.50	22.79	63.18	31.50	27.66
ICE [41]	ICCV 2021	20.54	57.50	20.39	10.24	29.81	69.41	38.35	34.32
JSIA-ReID [42]	AAAI 2020	38.10	80.7	36.90	–	43.80	86.2	52.90	–
OTLA [10]	ECCV 2022	29.9	–	27.1	–	29.8	–	38.8	–
ACCL [30]	CVPR 2023	57.27	92.48	51.78	34.96	56.23	90.19	62.74	58.13
ADCA [11]	MM 2022	45.51	85.29	42.73	28.29	50.60	89.66	59.11	55.17
H2H [43]	TIP 2021	30.15	65.92	29.40	–	–	–	–	–
QGM-BR(ours)	CMC 2025	58.36	93.27	52.92	40.85	57.02	90.87	63.98	57.74

As shown in Table 1, in the experiments on the SYSU-MM01 dataset, the proposed framework outperforms the best models on all evaluation metrics in both the full retrieval mode and indoor retrieval mode. Specifi-cally, in full retrieval mode, the model achieves a mAP of 52.92%, improving by 1.14% compared to the best model ACCL, and a Rank-1 accuracy of 58.36%, which is 1.09% higher than ACCL. In indoor retrieval mode, the mAP reaches 63.98%, an improvement of 1.01% over the best model AGW, and the Rank-1 accuracy is 57.02%, 0.79% higher than ACCL.

As shown in Table 2, in the experiments on the RegDB dataset, the proposed framework outperforms the best models in both test modes. In the Visible to Thermal mode, the mAP improves by 0.72%, and Rank-1 accuracy increases by 0.7% compared to the best model. In the Thermal to Visible mode, the mAP improves by 1.86%, and Rank-1 accuracy increases by 1.59%.

Table 2: Comparison of experimental results of different methods on the RegDB dataset (%)

Methods	Reference	Visible to thermal		Thermal to visible	
		Rank-1	mAP	Rank-1	mAP
SPCL [39]	NIPS 2020	13.59	14.68	11.70	13.56
AGW [14]	TPAMI 2021	70.05	66.37	70.49	65.90
MMT [40]	ICLR 2020	25.68	26.51	24.42	25.59
ICE [41]	ICCV 2021	12.98	15.64	12.18	14.82
JSIA-ReID [42]	AAAI 2020	48.50	49.30	48.10	48.90
OTLA [10]	ECCV 2022	32.90	29.70	32.10	28.60
ACCL [30]	CVPR 2023	69.48	65.41	69.85	65.17
ADCA [11]	MM 2022	67.20	64.05	64.48	63.81
H2H [43]	TIP 2021	23.81	18.87	–	–
QGM-BR (ours)	CMC 2025	70.75	67.09	72.08	67.76

In order to validate the model's excellent performance across diverse datasets, in addition to conducting experiments on the commonly used SYSU-MM01 and RegDB datasets, we also performed experiments on the newly released RGBNT201 dataset, with the results shown in Table 3. Since this dataset is newly introduced and unprocessed, it is not directly applicable to visible-infrared person re-identification, and

therefore, there are few research papers reporting results on it. We selected several methods that have shown strong performance in this context as competitors on this dataset.

Table 3: Comparison of experimental results of different methods on the RGBNT201 dataset (%)

Methods	Reference	Visible to thermal		Thermal to visible	
		Rank-1	mAP	Rank-1	mAP
TSLFN + HC [44]	Neurocomputing 2020	26.40	22.90	18.40	22.00
DDAG [45]	ECCV 2020	73.50	45.50	73.35	45.80
CM-NAS [46]	ICCV 2021	75.30	43.30	75.60	45.30
AGW [14]	TPAMI 2022	71.20	38.90	69.00	39.60
DTRM [47]	TIFS 2022	82.00	44.50	83.90	45.10
QGM-BR (ours)	CMC 2025	85.32	47.09	85.11	47.83

As shown in Table 3, in the experiments on the RGBNT201 dataset, the proposed framework outperformed the best model in both testing modes. In the Visible to Thermal mode, compared to the best model, the mAP improved by 3.32%, and the Rank-1 accuracy increased by 1.59%. In the Thermal to Visible mode, the mAP improved by 1.21%, and the Rank-1 accuracy increased by 2.03%. Overall, the proposed framework demonstrates high competitiveness across all three datasets.

4.5 Ablation Study

This section presents an ablation study to validate the effectiveness of each component of the proposed method. We use the DCL module described in Section 3.1 as the baseline and evaluate the performance after adding the QGM module and the BR module, as well as the effect of the contrastive loss function.

The experimental results are shown in Table 4. The addition of any single module among the QGM module, BR module, and contrastive loss function significantly improves the model's performance. Pairwise combinations of these three components also lead to notable performance improvements. When all three components are integrated together, they complement each other, achieving the best performance. Overall, each component positively contributes to the model's recognition performance, and their combined usage yields outstanding results. As shown in Fig. 6, we can also observe that the QGM and BR modules mutually complement each other, resulting in optimal performance. Notably, when the BR module operates independently without the QGM module, it treats each data point as an individual block due to the absence of QGM's matching results.

Table 4: Ablation experiments under the SYSU-MM01 dataset (%)

DCL	QGM	BR	$\mathcal{L}_{ns}$	All search			Indoor search		
				Rank-1	mAP	mINP	Rank-1	mAP	mINP
✓	✗	✗	✗	40.26	37.56	33.19	42.89	48.57	44.51
✓	✗	✓	✗	42.15	39.82	34.68	44.37	50.12	46.05
✓	✗	✗	✓	41.73	38.95	33.85	43.52	49.34	45.20
✓	✗	✓	✓	43.29	40.61	35.92	45.83	51.48	47.33
✓	✓	✗	✗	40.89	38.29	34.57	43.15	49.81	45.82

(Continued)

Table 4 (continued)

DCL	QGM	BR	$\mathcal{L}_{ns}$	All search			Indoor search		
				Rank-1	mAP	mINP	Rank-1	mAP	mINP
✓	✓	✗	✓	44.58	41.50	36.85	46.25	52.11	49.86
✓	✓	✓	✗	50.66	47.25	40.18	53.14	60.28	54.19
✓	✓	✓	✓	58.36	52.92	40.85	57.02	63.98	57.74

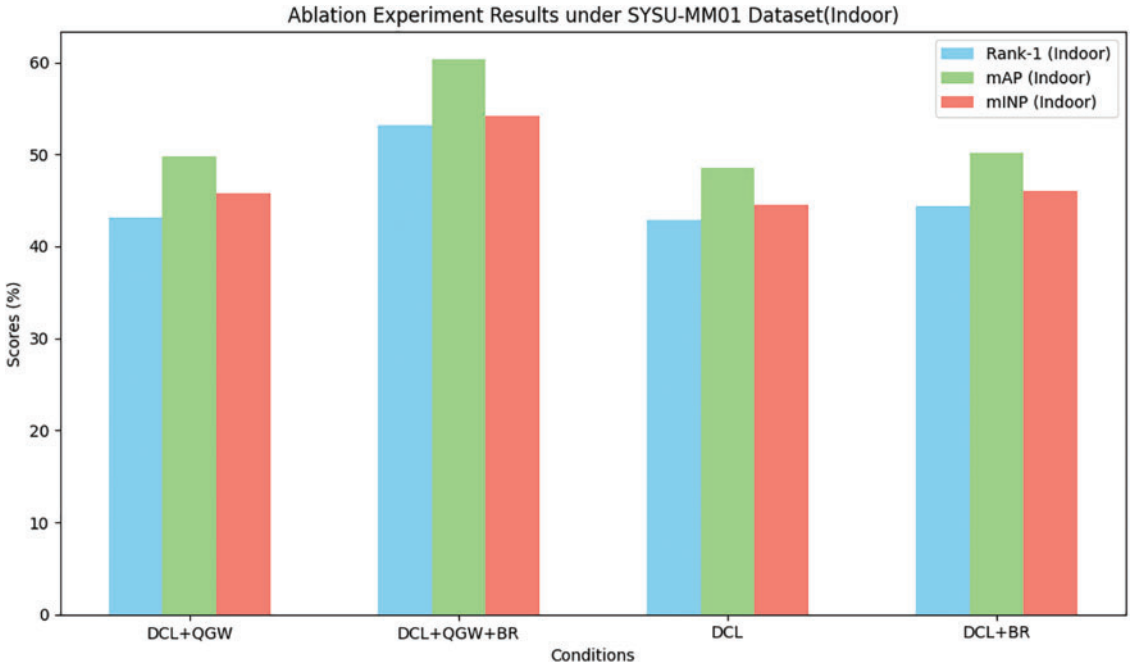

Figure 6: From the ablation experiment results under the perspective of the QGM and BR modules

5 Conclusion

We propose a framework based on Quadratic Graph Matching (QGM) and Block Reasoning (BR) to achieve reliable modality correspondence and efficient image updates. First, we transform the modality correspondence problem into a graph matching problem and use a quadratic matching strategy to effectively address the cluster imbalance issue. Additionally, we introduce the Block Reasoning module, which utilizes the affinity information between classes to enhance the precision of person search while simplifying the gallery update process. Extensive experiments demonstrate that the proposed method achieves state-of-the-art performance across multiple datasets. However, the affinity reasoning module is currently only applied during the testing phase, and the affinity information in the training phase has yet to be fully exploited. Future work will focus on integrating the affinity information with graph matching in the training phase for better performance.

Acknowledgement: The authors are grateful to all the editors and anonymous reviewers for their comments and suggestions, and thank all the members who have contributed to this work.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Junfeng Lin and Jialin Ma; data collection: Junfeng Lin and Hao Wang; analysis an interpretation of results: Weiguo Ding and Wei Chen; draft manuscript preparation: Mingyao Tang and Wei Chen. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Jin X, Lan C, Zeng W, Chen Z, Zhang L. Style normalization and restitution for generalizable person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 3140–9. doi:10.1109/cvpr42600.2020.00321.
2. Kalayeh MM, Basaran E, Gökmen M, Kamasak ME, Shah M. Human semantic parsing for person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 1062–71. doi:10.1109/CVPR.2018.00117.
3. Chen C, Ye M, Jiang D. Towards modality-agnostic person re-identification with descriptive query. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 15128–37. doi:10.1109/CVPR52729.2023.01452.
4. Kim M, Kim S, Park J, Park S, Sohn K. PartMix: regularization strategy to learn part discovery for visible-infrared person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 18621–32. doi:10.1109/CVPR52729.2023.01786.
5. Feng J, Wu A, Zheng WS. Shape-erased feature learning for visible-infrared person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 22752–61. doi:10.1109/CVPR52729.2023.02179.
6. Chen Y, Wan L, Li Z, Jing Q, Sun Z. Neural feature search for RGB-infrared person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 587–97. doi:10.1109/cvpr46437.2021.00065.
7. Wu Q, Dai P, Chen J, Lin CW, Wu Y, Huang F, et al. Discover cross-modality nuances for visible-infrared person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 4328–37. doi:10.1109/cvpr46437.2021.00431.
8. Lu Y, Wu Y, Liu B, Zhang T, Li B, Chu Q, et al. Cross-modality person re-identification with shared-specific feature transfer. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 13376–86. doi:10.1109/cvpr42600.2020.01339.
9. Zheng Z, Wang X, Zheng N, Yang Y. Parameter-efficient person re-identification in the 3D space. IEEE Trans Neural Netw Learn Syst. 2024;35(6):7534–47. doi:10.1109/tnnls.2022.3214834.
10. Wang J, Zhang Z, Chen M, et al. Optimal transport for label-efficient visible-infrared person re-identification. In: European Conference on Computer Vision; 2022; Tel Aviv, Israel. p. 93–109.
11. Yang B, Ye M, Chen J, Wu Z. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022; Lisboa, Portugal. p. 2843–51. doi:10.1145/3503161.3548198.
12. Huang H, Huang Y, Wang L. VI-diff: unpaired visible-infrared translation diffusion model for single modality labeled visible-infrared person re-identification. arXiv:2310.04122. 2023.
13. Cheng D, Huang X, Wang N, He L, Li Z, Gao X. Unsupervised visible-infrared person ReID by collaborative learning with neighbor-guided label refinement. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023; Ottawa ON, Canada. p. 7085–93. doi:10.1145/3581783.3612077.

14. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SCH. Deep learning for person re-identification: a survey and outlook. IEEE Trans Pattern Anal Mach Intell. 2022;44(6):2872–93. doi:10.1109/TPAMI.2021.3054775.
15. Huang X, Liu MY, Belongie S, Kautz J. Multimodal unsupervised image-to-image translation. In: Computer Vision—ECCV 2018; 2018; Munich, Germany. p. 179–96. doi:10.1007/978-3-030-01219-9_11.
16. Bai S, Ma B, Chang H, Huang R, Chen X. Salient-to-broad transition for video person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 7329–38. doi:10.1109/CVPR52688.2022.00719.
17. He W, Deng Y, Tang S, Chen Q, Xie Q, Wang Y, et al. Instruct-ReID: a multi-purpose person re-identification task with instructions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 17521–31. doi:10.1109/CVPR52733.2024.01659.
18. Li X, Xu Q, Zhang J, Zhang T, Yu Q, Sheng L, et al. Multi-modality affinity inference for weakly supervised 3D semantic segmentation. Proc AAAI Conf Artif Intell. 2024;38(4):3216–24. doi:10.1609/aaai.v38i4.28106.
19. Yao J, Anthony Q, Shafi A, Subramoni H, Panda DK, et al. Exploiting inter-layer expert affinity for accelerating mixture-of-experts model inference. In: 2024 IEEE International Parallel and Distributed Processing Symposium (IPDPS); 2024 May 27–31; San Francisco, CA, USA. p. 915–25. doi:10.1109/IPDPS57955.2024.00086.
20. Wang Z, Zhu F, Tang S, Zhao R, He L, Song J. Feature erasing and diffusion network for occluded person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 4744–53. doi:10.1109/CVPR52688.2022.00471.
21. Dai P, Ji R, Wang H, Wu Q, Huang Y. Cross-modality person re-identification with generative adversarial training. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI–18); 2018; Stockholm, Sweden. p. 677–83.
22. Yu X, Dong N, Zhu L, Peng H, Tao D. CLIP-driven semantic discovery network for visible-infrared person re-identification. IEEE Trans Multimed. 2025;1–13. doi:10.1109/TMM.2025.3535353.
23. Li H, Liu M, Hu Z, Nie F, Yu Z. Intermediary-guided bidirectional spatial-temporal aggregation network for video-based visible-infrared person re-identification. IEEE Trans Circuits Syst Video Technol. 2023;33(9):4962–72. doi:10.1109/TCSVT.2023.3246091.
24. Yao M, Wang H, Chen Y, Fu X. Between/within view information completing for tensorial incomplete multi-view clustering. IEEE Trans Multimed. 2024;27:1538–50. doi:10.1109/TMM.2024.3521771.
25. Wang H, Yao M, Chen Y, Xu Y, Liu H, Jia W, et al. Manifold-based incomplete multi-view clustering via bi-consistency guidance. IEEE Trans Multimed. 2024;26:10001–14. doi:10.1109/TMM.2024.3405650.
26. Xu Y, Lin L, Zheng WS, Liu X. Human re-identification by matching compositional template with cluster sampling. In: IEEE International Conference on Computer Vision; 2013 Dec 1–8; Sydney, NSW, Australia. p. 3152–9. doi:10.1109/ICCV.2013.391.
27. Zhang Z, Saligrama V. PRISM: person reidentification via structured matching. IEEE Trans Circuits Syst Video Technol. 2017;27(3):499–512. doi:10.1109/TCSVT.2016.2596159.
28. Rezatofighi SH, Milani A, Zhang Z, Shi Q, Dick A, Reid I. Joint probabilistic matching using m-best solutions. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 136–45. doi:10.1109/CVPR.2016.22.
29. Ye M, Ma AJ, Zheng L, Li J, Yuen PC. Dynamic label graph matching for unsupervised video re-identification. In: IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 5152–60. doi:10.1109/ICCV.2017.550.
30. Wu Z, Ye M. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 9548–58. doi:10.1109/CVPR52729.2023.00921.
31. Huang Z, Liu J, Li L, Zheng K, Zha ZJ. Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification. Proc AAAI Conf Artif Intell. 2022;36(1):1034–42. doi:10.1609/aaai.v36i1.19987.
32. Liu J, Sun Y, Zhu F, Pei H, Yang Y, Li W. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 19344–53. doi:10.1109/CVPR52688.2022.01876.

33. Jia M, Zhai Y, Lu S, Ma S, Zhang J. A similarity inference metric for RGB-infrared cross-modality person re-identification. arXiv:2007.01504. 2020.
34. Fang X, Yang Y, Fu Y. Visible-infrared person re-identification via semantic alignment and affinity inference. In: IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 11270–9. doi:10.1109/ICCV51070.2023.01035.
35. Wu A, Zheng WS, Yu HX, Gong S, Lai J. RGB-infrared cross-modality person re-identification. In: IEEE International Conference on Computer Vision (ICCV); 2017; Venice, Italy. p. 5390–9. doi:10.1109/ICCV.2017.575.
36. Nguyen DT, Hong HG, Kim KW, Park KR. Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors. 2017;17(3):605. doi:10.3390/s17030605.
37. Zheng A, Wang Z, Chen Z, Li C, Tang J. Robust multi-modality person re-identification. Proc AAAI Conf Artif Intell. 2021;35(4):3529–37. doi:10.1609/aaai.v35i4.16467.
38. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi:10.1109/CVPR.2016.90.
39. Ge Y, Zhu F, Chen D, Zhao R, Li H. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. Adv Neural Inf Process Syst. 2020;33:11309–21.
40. Ge Y, Chen D, Li H. Mutual mean-teaching: pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv:2001.01526. 2020.
41. Chen H, Lagadec B, Bremond F. ICE: inter-instance contrastive encoding for unsupervised person re-identification. In: IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 14940–9. doi:10.1109/ICCV48922.2021.01469.
42. Wang GA, Zhang T, Yang Y, Cheng J, Chang J, Liang X, et al. Cross-modality paired-images generation for RGB-infrared person re-identification. Proc AAAI Conf Artif Intell. 2020;34(7):12144–51. doi:10.1609/aaai.v34i07.6894.
43. Liang W, Wang G, Lai J, Xie X. Homogeneous-to-heterogeneous: Unsupervised learning for RGB-infrared person re-identification. IEEE Trans Image Process. 2021;30:6392–407. doi:10.1109/TIP.2021.3092578.
44. Zhu Y, Yang Z, Wang L, Zhao S, Hu X, Tao D. Hetero-center loss for cross-modality person re-identification. Neurocomputing. 2020;386:97–109. doi:10.1016/j.neucom.2019.12.100.
45. Ye M, Shen J, Crandall J, Shao D, Luo L, J et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Computer Vision-ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. p. 229–47.
46. Fu C, Hu Y, Wu X, Shi H, Mei T, He R. CM-NAS: cross-modality neural architecture search for visible-infrared person re-identification. In: IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 11803–12. doi:10.1109/ICCV48922.2021.01161.
47. Ye M, Chen C, Shen J, Shao L. Dynamic tri-level relation mining with attentive graph for visible infrared re-identification. IEEE Trans Inf Forensics Secur. 2021;17:386–98. doi:10.1109/TIFS.2021.3139224.