Reliable Multi-Modal Object Re-Identification via Modality-Aware Graph Reasoning

Xixi Wan, Aihua Zheng, Zi Wang, Bo Jiang, Jin Tang, and Jixin Ma

***Abstract*—Multi-modal data provides abundant and diverse object information, crucial for effective modal interactions in the Re-Identification (ReID) task. However, existing approaches often overlook the quality variations in local features and fail to fully leverage the complementary information across modalities, particularly in cases where features are of low quality. In this paper, we propose to address this issue by leveraging a novel graph reasoning model, termed the Modality-aware Graph Reasoning Network (MGRNet). Specifically, we first construct modality-aware graphs to enhance the extraction of fine-grained local details by effectively capturing and modeling the relationships between patches. Subsequently, the selective graph nodes swap operation is employed to alleviate the adverse effects of low-quality local features by considering both local and global information, enhancing the representation of discriminative information. Finally, the swapped modality-aware graphs are fed into the local-aware graph reasoning module, which propagates multi-modal information to yield a reliable feature representation. Another advantage of the proposed graph reasoning approach is its ability to reconstruct missing modal information by exploiting inherent structural relationships, thereby minimizing disparities between different modalities. Experimental results on four benchmarks (RGBNT201, Market1501-MM, RGBNT100, MSVR310) indicate that the proposed method achieves state-of-the-art performance in multi-modal object ReID. Our code is available at https://github.com/wanxixi11/MGRNet**

***Index Terms*—Multi-modal object re-identification, modality-aware graph, selective graph nodes swap, graph reasoning network, modality missing.**

Received 19 October 2025; revised 16 April 2026; accepted 16 May 2026. Date of publication 25 May 2026; date of current version 3 June 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62372003; in part by the Natural Science Foundation of Anhui Province under Grant 2308085Y40 and Grant 2408085J037; in part by the Key Technologies Research and Devlopment Program of Anhui Province under Grant 202423k09020039; and in part by the Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, under Grant 2024A004. The associate editor coordinating the review of this article and approving it for publication was Dr. Daniel Moreira. *(Corresponding authors: Aihua Zheng; Bo Jiang.)*

Xixi Wan and Aihua Zheng are with the School of Artificial Intelligence and the State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology, Anhui University, Hefei 230601, China, and also with Anhui Provincial Key Laboratory of Intelligent Detection and Diagnosis for Traffic Infrastructure, Hefei 230032, China (e-mail: xixiwan11@163.com; ahzheng214@foxmail.com).

Zi Wang is with the School of Biomedical Engineering, Anhui Medical University, Hefei 230032, China, and also with the Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei 230601, China (e-mail: ziwang1121@foxmail.com).

Bo Jiang and Jin Tang are with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: jiangbo@ahu.edu.cn; tangjin@ahu.edu.cn).

Jixin Ma is with the School of Computing and Mathematical Sciences, University of Greenwich, SE10 9LS London, U.K. (e-mail: j.ma@greenwich.ac.uk).

Digital Object Identifier 10.1109/TIFS.2026.3696569

I. Introduction

RECENTLY, multi-modal data has gained prominence as a promising direction in computer vision, particularly for the task of object Re-Identification (ReID) [1], [2], [3]. Compared with single-modal and cross-modal data [4], [5], [6], [7], [8], [9], [10], multi-modal data can capture object features more comprehensively by integrating information from different data sources, which is beneficial for practical object ReID scenarios. For example, Near Infrared (NIR) images can provide more visible information especially in low illumination, while Thermal Infrared (TIR) images have a strong ability to penetrate haze and smog [11], [12], [13], [14], [15], [16]. Previous works have demonstrated the effectiveness of multiple modalities in enhancing the performance of the object ReID task [12], [14], [17], [18], [19], [20], [21], [22], [23], [24], [25]. However, we observe that existing multi-modal methods often rely on global feature extraction to exploit the complementary information between different modalities [17], [20], [21], [22], which ignores mining fine-grained local cues for the multi-modal ReID task. To address this issue, part-based methods are introduced to multi-modal object ReID by simply dividing features either randomly or based on object part positions inferred from vision encoders [3], [14], [25], [26]. *Although these part-based methods can extract and utilize local information, they fail to consider the quality differences between local features of different modalities, which are ineffective for low-quality local features, as shown in Fig. 1 (a).* Additionally, in real-world scenarios, the issue of missing modalities is common and often unavoidable. For example, Zhu et al. [27] and Wang et al. [13], [23] propose image and feature reconstruction to compensate for the absence of images, respectively. Wang et al. [28] utilize zero padding to reconstruct information. *However, existing works mainly focus on feature- or image-level reconstruction while neglecting the essential structural relationships and dependencies among features of the missing modality. This leads to the generation of missing modalities with inconsistent structure and semantics, impacting the model's performance.*

To address the above limitations, we propose a new Modality-aware Graph Reasoning Network (MGRNet) that can effectively boost information interactions and recover missing modalities for multi-modal object ReID. Specifically, the modality-aware graphs are first utilized to facilitate the extraction of important local details by modeling the relationships between patches for multi-modal data. Obviously, existing low-quality information somewhat increases the difficulty of learning spectral image features. Meanwhile, current approaches often struggle to fully obtain the comple-

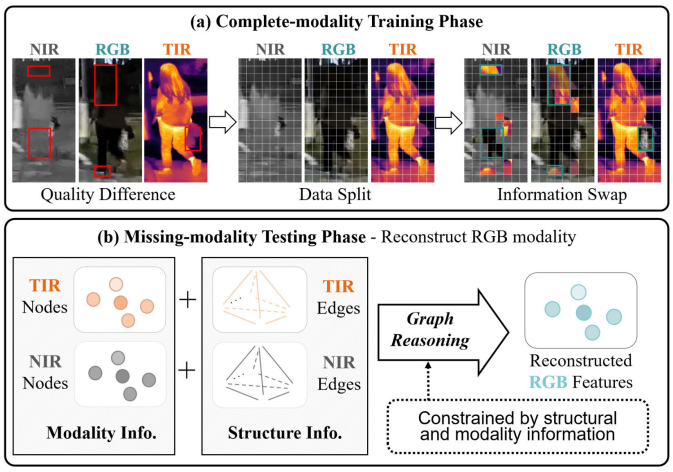

Fig. 1. (a) Due to quality differences in local features among modalities, we first split the data to obtain more detailed local information, and then perform an information swap (achieved by GRMI described in Sec III-C). (b) When RGB is missing in the testing phase, we leverage graph reasoning trained with constraints from modality and structural information to restore features, combining existing NIR and TIR features (nodes) and their relationships (edges), achieved by GRMM described in Sec III-E.

mentarity and dependencies between modalities in Fig. 1 (a). These problems prompt us to propose a new multi-modal fusion method for object ReID via the Selective Graph Nodes Swap (SGNS) operation. SGNS first identifies low-quality node pairs with relatively small edge weights in the graph, as these nodes usually correspond to local regions that are weakly related to their neighboring nodes. The selection is then further refined based on the similarity between local and global tokens. As a result, the identified low-quality nodes are selected by both weak structural connectivity and low semantic consistency with the overall information. Thus, this operation is designed to effectively mitigate the impact of low-quality local features by considering both local and global information, enhancing the discriminative information. As shown in Fig. 1 (a), swapping local features addresses issues related to low-quality features, thus improving the effectiveness of modal representations. Subsequently, we feed the swapped modality-aware graphs into the local-aware graph reasoning module to achieve multi-modal information propagation, thus yielding reliable feature representations. Another key aspect of the proposed graph reasoning is that it can recover the features of missing modalities by exploiting their structural relationships, thereby minimizing the undesired disparity between different modalities. This recovered loss is intended to reduce the gap between the reconstructed and real representations, as well as to minimize the disparity between modalities. During the testing phase, the reconstruction module can be directly utilized to generate features for the missing modality, as shown in Fig. 1 (b).

Overall, MGRNet can capture rich, fine-grained local features while reducing the influence of low-quality local tokens by considering both local and global information, jointly promoting information interaction for objects. To the best of our knowledge, this is the first attempt to leverage graph reasoning for both modal interaction and missing problems in multi-modal object ReID, which achieves outstanding performance on common object datasets. The main contributions of this work are summarized as,

- We propose the Modality-aware Graph Reasoning Network (MGRNet) to jointly capture the crucial structure relationships and complementary information among different modal tokens, solving both modal interaction and modal missing problems for multi-modal object ReID.
- We introduce multiple modality-aware graphs to incorporate structural information of local features and design the selective graph nodes swap operation to effectively alleviate the impact of low-quality local features.
- The MGRNet inherently possesses the ability to reconstruct the features of missing modalities based on their structural relationships, as well as to minimize the disparity between modalities.
- Extensive experiments on four public benchmarks show that MGRNet achieves superior performance over state-of-the-art approaches for multi-modal object ReID, validating its reliability and effectiveness.

II. Related Works

A. Multi-Modal Object ReID

At present, existing research has made significant progress in multi-modal object ReID [12], [13], [23], [25], [29]. For example, in works [12], [23], [25], [29], they utilize and fuse multiple source data to provide a more comprehensive interaction of modalities. Among that, some methods leveraging global-based methods have been proposed [21], [22], [24]. CCNet [22] proposes CDC loss to solve discrepancies in both modalities and samples, generating discriminative multi-modal feature representations for vehicle ReID. HTT [24] proposes to utilize CIM loss to enlarge the differentiation between distinct identity samples and then apply MTT to optimize the generalization capabilities of the model. DeMo [13] first decouples multi-modal features to obtain modal-specific and -shared information. Then, an Attention-Triggered Mixture of Experts (ATMoE) is leveraged to enhance modal interactions via attention-guided operation. However, the above methods only focus on the complementarity and dependencies of global information while overlooking the deep interaction of local information. For multi-modal object ReID, it is well-known that local features provide rich, fine-grained object information, which is essential for effective feature interaction and reconstruction. Thus, part-based multi-modal methods are proposed [12], [19], [23], [25]. PFNet [12] proposes a progressive fusion network to effectively fuse different spectral features, including RGB, NIR, and TIR, for the object ReID task. IEEE [19] proposes a relation-based embedding module to embed the global information into fine-grained local features, boosting feature representations for multi-modal ReID. EDITOR [25] introduces object-centric tokens for multi-modal ReID. This method proposes SFTS and HMA to select and aggregate multi-modal tokenized features. TOP-ReID [23] proposes a Token Permutation Module (TPM). This module can align multi-spectral images and utilize the current global

token to perceive the local token of other modalities. IDEA [14] adaptively generates sampling positions by aggregating multi-modal information to facilitate the interaction between global features and local information.

Although these methods can extract and use local information, they either adopt random segmentation or divide features into multiple parts based on object parts, ignoring the quality difference between local features of different modalities, thereby creating certain noise for the extracted features.

B. Graph-Based Object ReID

Graph-based methods have been widely applied in object ReID, which can model relationships between single-modal images and promote feature representations [30], [31], [32], [33], [34]. For instance, Nguyen et al. [32] propose a graph-based person signature, fusing detailed person descriptions and visual features into a graph. Jiang et al. [35] propose PH-GCN to tackle the single-modal person ReID problem, offering a unified solution that effectively integrates local, global, and structural feature representations. He et al. [36] propose the PGGANet method, utilizing a self-adaptive graph attention convolution to learn the contribution matrix of local information. Sun et al. [2] propose a polymorphic masked wavelet graph convolutional network to disentangle content and degradation features of cross-modality images. Lv et al. [37] propose a novel edge weight-embedding graph convolutional network that embeds human joints and bones into the feature representation of object ReID. While these methods achieve relatively good results in single-modal object ReID, they mainly emphasize interactions for global and local features, neglecting the quality difference of local features and the complementary information across modalities.

Notably, unlike the above methods, our MGRNet is the first to propose utilizing graph reasoning to exploit both modal interaction and missing problems by considering both global and local features in multi-modal object ReID. We first design a Graph Reasoning on Modal Interaction (GRMI) strategy to adaptively learn information between patches of multi-modal images, meanwhile presenting a novel Graph Reasoning on Missing Modality (GRMM) strategy to compensate for missing modal information and reduce the difference between modalities.

III. The Proposed Method

A. Overview

In this section, we present a novel method to enhance information interaction and reconstruct missing modality, called Modality-aware Graph Reasoning Network (MGRNet) for multi-modal object ReID. This method consists of four main parts: (1) Initial Feature Extraction, (2) Graph Reasoning on Modal Interaction, (3) Global-aware Multi-Head Attention, and (4) Graph Reasoning on Missing Modality, as illustrated in Fig. 2. Finally, a comprehensive training loss is utilized to optimize the entire proposed method. Below, we will introduce the main parts in detail.

B. Initial Feature Extraction

We adopt multi-branch backbones for extracting feature representations from multi-modal data, given their excellent performance for multi-modal images in object ReID [14], [24], [25]. To preserve special information of each modality and extract better initial features of different modalities, the multi-branch backbones are non-shared. Formally, this process can be defined as follows,

$$X^m = \{X_g^m, X_l^m\} = \xi^m(I^m), \tag{1}$$

where ξ^m denotes the vision encoder (ViT or CLIP) [14], [23], [24]. I^m is the input image for the m-th modality, with $m \in \{N, R, T\}$ where N, R, and T refer to the NIR, RGB, and TIR modalities, respectively. $X^m \in \mathbb{R}^{(P+1)\times D}$ is the extracted features for the m-th modality. $X_l^m \in \mathbb{R}^{P\times D}$ and $X_g^m \in \mathbb{R}^D$ represent patch tokens and class tokens of input images, respectively. $P = 128$ is the number of local tokens and D denotes the obtained dimensions of embedding tokens.

C. Graph Reasoning on Modal Interaction

Graph Reasoning on Modal Interaction (GRMI) focuses on the quality inconsistency of local features for multi-modal data. This interaction improves fine-grained representation learning through a constructed node-selective graph and local-aware graph reasoning. Next, we describe GRMI in detail.

1) Modality-Aware Graph Learning (MGL): To fully encode the local features obtained above of multi-modal information, we build a modality-aware graph $G(V^m, E^m)$ for each modality. For the m-th modality, the nodes V^m of our modality-aware graph are represented as the set of local features of images. To be specific, let $X_l^m = \{X_{l_1}^m, X_{l_2}^m \cdots X_{l_P}^m\} \in \mathbb{R}^{P\times D}$ denote the collection features of all patches in the m-th modality. The edges E^m of our modality-aware graph connect different patches, utilizing an adjacency matrix A^m to represent structural relationships among local features. It is learned dynamically by computing the Euclidean distance [38] to construct relationships as,

$$A_{l_{ij}}^m = 1 - \sigma\big((\psi(X_{l_i}^m, X_{l_j}^m) + \alpha) \times \beta\big), \tag{2}$$

where $X_{l_i}^m$ and $X_{l_j}^m$ represent the feature vectors of the i-th and j-th patches for the m-th modality, respectively. And ψ is the Euclidean distance [38]. α and β are two learnable parameters, and σ is the Sigmoid activation function.

2) Selective Graph Nodes Swap (SGNS): To further promote interaction among obtained local features, the Selective Graph Nodes Swap (SGNS) operation is designed to reduce the impact of low-quality local details between modalities through considering both local and global information [39], [40], [41]. As shown in Fig. 3, this operation consists of two main steps. **Firstly**, the Top-k method is employed to select k small values for edges of each graph and then find the poor patches corresponding. **Secondly**, based on the above method, we also introduce global tokens to select more precise local nodes about low-quality patches. Specifically, the process first calculates the similarity between the global feature and each local feature as follows,

$$W^m = 1 - \phi\big(\psi(X_g^m, X_l^m)\big), \tag{3}$$

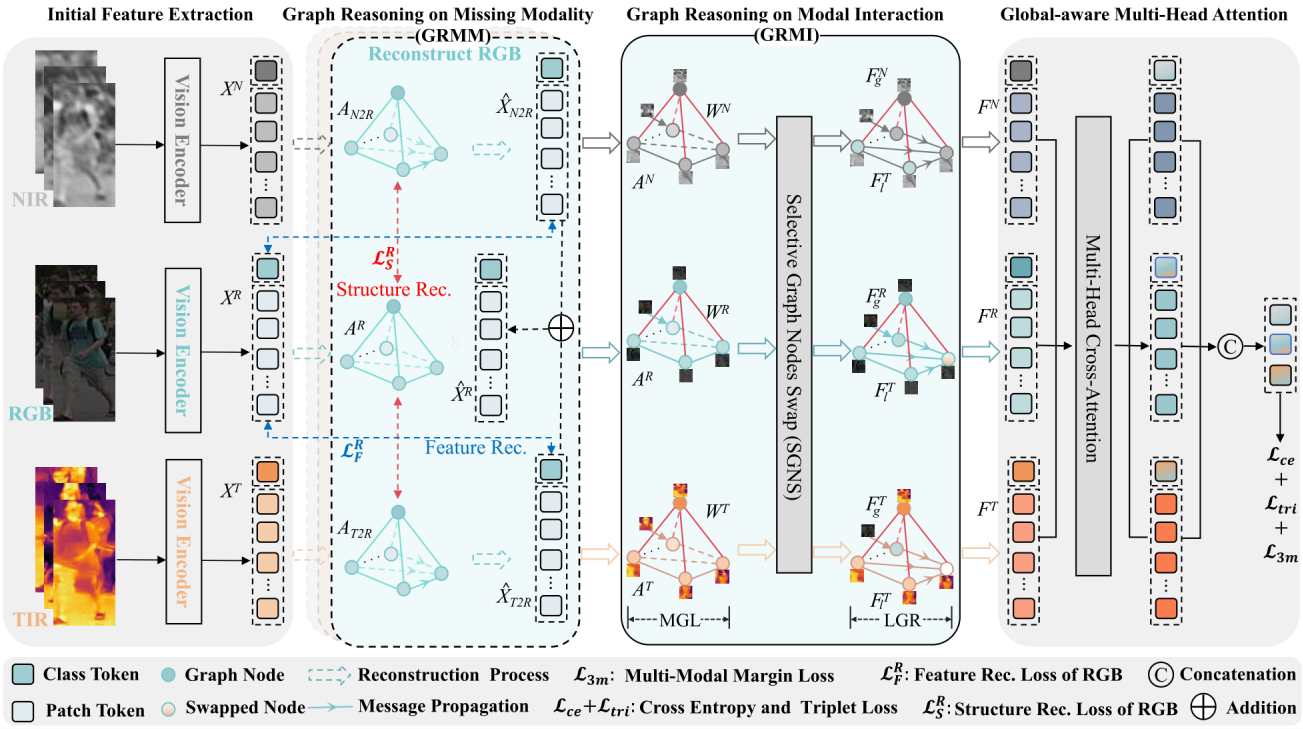

Fig. 2. The overall network structure of the proposed MGRNet. For complete multi-modal training and testing, initial feature extraction first employs the multi-branch vision encoders on the multi-modal images to obtain the initial features. Secondly, graph reasoning on modal interaction is employed to alleviate low-quality tokens of each modality. Finally, the enhanced features are generated with global-aware multi-head attention and the fused features are fed into the classifiers to get the ReID results. Furthermore, graph reasoning on the missing modality strategy is designed to restore features based on their structural relationships for the missing modality problems.

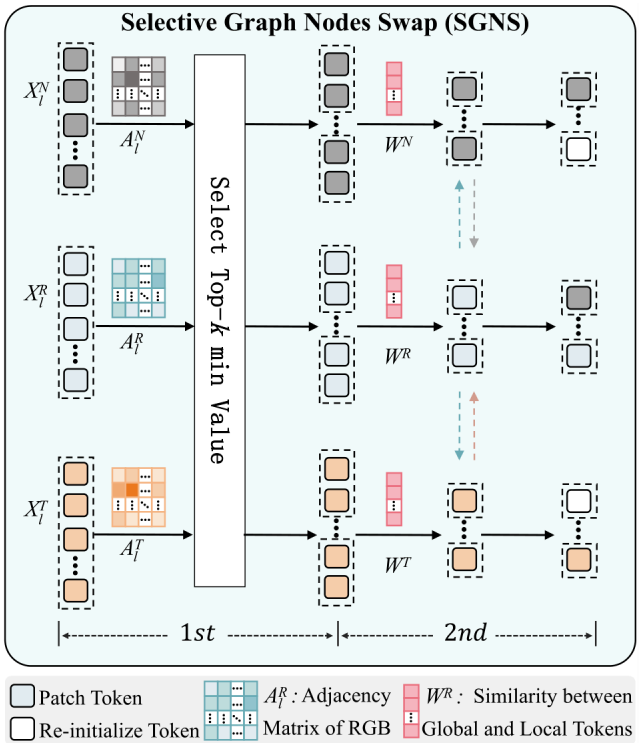

Fig. 3. The process of multiplying selective graph nodes swap by considering both local and global information.

where $W^m = \{w_1^m, w_2^m \cdots w_P^m\} \in \mathbb{R}^P$ denotes the strength of the relationship between global and local features. ϕ is the Softmax of the non-linear activation function. And then W^m is applied to screen the first step to get the poor nodes, which is beneficial for obtaining lower-quality patches.

As we know, local features perform differently across modalities, resulting in varying quality. Thus, local patches between RGB, NIR, and TIR are exchanged based on their respective quality; for example, the poor nodes of RGB will be replaced by the mean features of NIR and TIR. Finally, we obtain the updated node features of local cues.

Note that we replace the inferior patches by leveraging better patches of other modalities while taking into account the correlation between patches. So the updated $X_{l_i}^m$ is defined as,

$$X_{l_i}^R = Swap\left(X_{l_i}^R, \frac{1}{2}(X_{l_i}^N + X_{l_i}^T)\right), \tag{4}$$

$$X_{l_i}^N = Swap\left(X_{l_i}^N, X_{l_i}^R\right), X_{l_i}^T = Swap\left(X_{l_i}^T, X_{l_i}^R\right). \tag{5}$$

In addition, to avoid the exchange of lower-quality patches, we also design a mechanism to determine whether the swapped patch is a poor patch. If so, we re-initialize the current patch token that is set to an all-zero matrix and then learn the feature representation of the current patch through the local neighbors of the current patch; otherwise, we exchange this patch as Fig. 3. This design can not only facilitate the learning of information from other modalities but also mitigate the impact of low-quality local features by considering both local and global information.

3) Local-Aware Graph Reasoning (LGR): Based on the above SGNS operation, we obtain enhanced node representations of each modality-aware graph. Accordingly, the multi-layer local-aware graph reasoning module is employed

to learn local patches representation of multi-modal data for better quality node patches. The message propagation rule is defined as follows,

$$F_l^{(m,\tilde{l}+1)} = \delta(A_l^m F_l^{(m,\tilde{l})} \Theta^{(m,\tilde{l})}), \tag{6}$$

where $\tilde{l} = 0, 1 \cdots \tilde{L}-1$ denotes the $\tilde{l}$ layer of GCN and $F_l^{(m,0)} = X_l^m$. $\Theta^{(m,\tilde{l})}$ is the learnable transformation matrix and $F_l^{(m,\tilde{L})}$ is briefly represented as F_l^m. δ denotes the non-linear activation function ReLU.

D. Global-Aware Multi-Head Attention

Let $F_l^m = \{F_{l_1}^m, F_{l_2}^m \cdots F_{l_P}^m\}$ denote the obtained patches via the above graph reasoning on the modal interaction strategy. We introduce the global-aware multi-head attention module, which can capture richer node representations by increasing the interaction of global and local information. Formally, we apply a linear projection layer to obtain a query matrix $Q^m \in \mathbb{R}^D$ for global token $F_g^m = X_g^m$ and different linear project layer to get key matrix $K^m \in \mathbb{R}^{P\times D}$ and value matrix $V^m \in \mathbb{R}^{P\times D}$ for local token F_l^m. Then, multi-head attention with H heads is utilized to aggregate local information from different patches into global feature representations. This interaction operation is defined in the h-th head as follows,

$$\hat{F}^{(m,h)} = \phi\left[\frac{Q^{(m,h)}(K^{(m,h)})^T}{\tau}\right] V^{(m,h)}, \tag{7}$$

where τ is the scale factor and ϕ is the Softmax of the non-linear activation function. $\hat{\mathbf{F}}^{(m,h)}$ represents obtained output feature of the h-th head for the m-th modality. To aggregate the information from all heads H, we utilize a concatenation operation to get a new class token as,

$$\hat{\mathbf{F}}^m = \mathrm{Con}\big(\hat{F}^{(m,1)} \cdots \hat{F}^{(m,H)}\big). \tag{8}$$

Finally, we concatenate features of all modalities to obtain the fused features $\mathbf{Z}$ as follows,

$$\mathbf{Z} = \mathrm{Con}(\hat{\mathbf{F}}^N, \hat{\mathbf{F}}^R, \hat{\mathbf{F}}^T). \tag{9}$$

E. Graph Reasoning on Missing Modality

To address the modality-missing issue in the real world, the MGRNet also designs the Graph Reasoning on Missing Modality (GRMM) strategy. This strategy compensates for missing modal information by leveraging both the available modality features and their structural relations. As shown in Fig. 2, GRMM first applies feature reconstruction to enhance feature representations, then leverages structure reconstruction to learn the relationships between corresponding tokens [42], [43]. To be more specific, assuming that the RGB modality is missing, we leverage the existing NIR and TIR modalities to recover the tokens of the missing RGB modality. We first compute the original structure relationships between all real tokens that include local and global features of each modality, as follows,

$$A_{ij}^m = 1 - \sigma(D_{ij}^m), D_{ij}^m = \psi(X_i^m, X_j^m)), \tag{10}$$

where X_i^m and X_j^m represent the feature vectors of the tokens i and j respectively for the m-th modality. ψ is the Euclidean distance [38]. σ is the Sigmoid activation function. Next, we construct dynamically reconstructed structure relationships A_{N2R}, $A_{T2R} \in \mathbb{R}^{(P+1)\times(P+1)}$ as follows,

$$A_{N2R} = 1 - \sigma(t_N D^N), A_{T2R} = 1 - \sigma(t_T D^T), \tag{11}$$

where t_N and t_T are two learnable hyperparameters. And then equipped with a layer-wise GCN to propagate messages for recovering feature $\hat{X}_{N2R}$ and $\hat{X}_{T2R}$. This recovered process is defined as,

$$\hat{X}_{N2R}^{\hat{l}+1} = \delta(A_{N2R}\hat{X}_{N2R}^{\hat{l}}\Theta^{(N,\hat{l})}), \tag{12}$$

$$\hat{X}_{T2R}^{\hat{l}+1} = \delta(A_{T2R}\hat{X}_{T2R}^{\hat{l}}\Theta^{(T,\hat{l})}), \tag{13}$$

where $\hat{l} = 0, 1 \cdots \hat{L}-1$ denotes the $\hat{l}$ layer of GCN and $\hat{X}_{N2R}^0 = X^N, \hat{X}_{T2R}^0 = X^T \in \mathbb{R}^{(P+1)\times D}$. $\Theta^{(N,\hat{l})}$ and $\Theta^{(T,\hat{l})}$ are the layer-wise trainable transformation matrices. The outputs via $\hat{L}$ layer-wise GCN are $\hat{X}_{N2R}^{\hat{L}}, \hat{X}_{T2R}^{\hat{L}}$ which are briefly denoted as $\hat{X}_{N2R}, \hat{X}_{T2R}$. In the absence of the RGB modality, we dynamically generate its token features using the information from NIR and TIR modalities as follows,

$$\hat{X}^R = \frac{1}{2}\big(\hat{X}_{N2R} + \hat{X}_{T2R}\big). \tag{14}$$

Then, the reconstructed token features $\hat{X}^R$, X^N, X^T are fed into the proposed graph reasoning on the modal interaction strategy to jointly fuse multi-modal data for ReID. The proposed GRMM strategy not only solves the limitation of the receptive field of CNN architecture by convolution on the graph structure but also effectively tackles the challenge of missing modality in the ReID task.

F. Training Loss

To promote the reconstruction ability, the recovered features are constrained via leveraging both feature reconstruction loss $\mathcal{L}_F^R$ and structure reconstruction loss $\mathcal{L}_S^R$ as follows,

$$\mathcal{L}_F^R = \frac{1}{P+1}\sum_{p=1}^{P+1}\|\hat{X}_{N2R} - X^R\|^2 + \|\hat{X}_{T2R} - X^R\|^2, \tag{15}$$

$$\mathcal{L}_S^R = \frac{1}{P+1}\sum_{p=1}^{P+1}\|A_{N2R} - A^R\|^2 + \|A_{T2R} - A^R\|^2, \tag{16}$$

$$\mathcal{L}^R = \mathcal{L}_F^R + \mathcal{L}_S^R. \tag{17}$$

We observe that the constraints can also reduce the gap between the generated RGB modality and other modalities. Similarly, if other modalities are missing, reconstruction can also be achieved to obtain a smaller modal distribution gap. Thus, the proposed GRMM strategy is trained under the constraints of multi-modal reconstruction loss $\mathcal{L}_{MR}$ as,

$$\mathcal{L}_{MR} = \mathcal{L}^R + \mathcal{L}^N + \mathcal{L}^T. \tag{18}$$

In the training phase, our training loss comprises several components, including Vision Encoders, Graph Reasoning on Missing Modality, Graph Reasoning on Modal Interaction, and Global-aware Multi-Head Attention. We optimize the entire network by utilizing the above multi-modal reconstruction

TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS ON THE PERSON REID DATASETS (IN %). THE SYMBOL * INDICATES VIT-BASED METHODS, WHILE † REPRESENTS CLIP-BASED METHODS FOR OUR PROPOSED MGRNET. THE BEST AND SECOND BEST RESULTS ARE MARKED IN BOLD AND UNDERLINE, RESPECTIVELY

	Methods	Publication	Structure	RGBNT201				Market1501-MM			
				mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Single	HACNN [46]	CVPR18	CNN	21.3	19.0	34.1	42.8	42.9	69.1	86.6	92.2
	MLFN [47]	CVPR18	CNN	26.1	24.2	35.9	44.1	42.7	68.1	87.1	92.0
	OSNet [7]	ICCV19	CNN	25.4	22.3	35.1	44.7	39.7	69.3	86.7	91.3
	TransReID [9]	ICCV21	ViT	63.8	65.8	78.5	83.9	73.0	88.9	95.8	97.6
Multi	HAMNet [17]	AAAI20	CNN	27.7	26.3	41.5	51.7	60.0	82.8	92.5	95.0
	PFNet [12]	AAAI21	CNN	38.5	38.9	52.0	58.4	60.9	83.6	92.8	95.5
	IEEE [19]	AAAI22	CNN	46.4	47.1	58.5	64.2	64.3	83.9	93.0	95.7
	UniCat [20]	NIPSW23	ViT	57.0	55.7	-	-	-	-	-	-
	EDITOR [25]	CVPR24	ViT	65.7	68.8	82.5	89.1	77.4	90.8	96.8	98.3
	RSCNet [26]	TCSVT24	ViT	68.2	72.5	-	-	-	-	-	-
	HTT [24]	AAAI24	ViT	71.1	73.4	83.1	87.3	67.2	81.5	95.8	97.8
	TOP-ReID [23]	AAAI24	ViT	72.3	76.6	84.7	89.4	82.0	92.4	97.6	98.6
	DeMo [13]	AAAI25	ViT	73.7	80.5	88.3	91.5	78.0	90.7	96.8	98.2
	MGRNet*	**Ours**	ViT	**78.4**	**82.5**	**90.9**	**95.2**	**84.4**	**94.0**	**98.2**	**98.9**
	MambaPro [3]	AAAI25	CLIP	78.9	83.4	89.8	91.9	84.1	92.8	97.7	98.7
	DeMo [13]	AAAI25	CLIP	79.0	82.3	88.8	92.0	83.6	93.1	97.5	98.7
	IDEA [14]	CVPR25	CLIP	80.2	82.1	90.0	**93.3**	-	-	-	-
	MGRNet†	**Ours**	CLIP	**80.5**	**85.0**	**90.0**	92.6	**84.6**	**93.6**	**97.7**	**98.8**

loss, multi-modal margin [19], cross entropy loss [44] and triplet loss [45], minimizing the sum of all losses as follows,

$$\mathcal{L} = \mathcal{L}_{MR} + \mathcal{L}_{ce} + \mathcal{L}_{tri} + \mathcal{L}_{3m}. \tag{19}$$

$\mathcal{L}_{3m}$ is removed for the proposed MGRNet of CLIP-based vision encoders. This overall network is optimized in an end-to-end manner.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we evaluate the effectiveness of the proposed MGRNet on four commonly used datasets and compare it with some other related works.

A. Dataset and Evaluation Metrics

We utilize four commonly used multi-modal ReID datasets to evaluate our MGRNet, including two-person ReID datasets (RGBNT201 [12], Market1501-MM [19]) and two-vehicle ReID datasets (RGBNT100 [17] and MSVR310 [22]). RGBNT201 [12] has 201 identities with four different viewpoints, with varying lighting and background complexity challenges. Market1501-MM [19] is a virtual multi-modal dataset generated from a single-modality dataset [48] via the cycle-GAN method [27] for person ReID, yielding 1,501 identities with reducing 60% of the brightness. RGBNT100 [17] is a large-scale dataset, comprising 17,250 image triples of 100 vehicles across RGB, NIR, and TIR modalities. In contrast, MSVR310 [22] is a smaller-scale dataset that includes more complex scenarios. For evaluation metrics, we adopt the previous evaluation strategy, utilizing mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-K (K = 1, 5, 10) for all used datasets. Higher values of these metrics imply better model performance.

B. Implementation Details

The proposed model is implemented using PyTorch with one RTX 4090 GPU. For all used datasets, we resize images into $256 \times 128 \times 3$ pixels for person datasets and $128 \times 256 \times 3$ pixels for vehicle datasets. Furthermore, we adopt previous random erasure, random flipping, and padding in the training [9]. The graph is constructed with 128 local nodes and 1 global node. For the selective graph node swap strategy, we set the Top-k parameter to 20. As shown in the later experiments, this setting provides excellent performance. Additionally, we adopt pre-trained ViT/CLIP as the vision encoder. The ViT-based and CLIP-based MGRNet are trained for 80 and 40 epochs, respectively. SGD and Adam [59] are used as the optimizers for the two variants, respectively, with a momentum of 0.9 and a weight decay of 0.0001. The initial learning rates are set to 0.0066 and 0.00035, respectively, and are scheduled using a warm-up strategy followed by cosine decay. During training, the batch size is set to 64, consisting of 4 identities with 16 images per identity. The number of warm-up iterations is set to 10 for all datasets.

C. Comparisons With State-of-the-Art Methods

We conduct comprehensive comparisons with state-of-the-art methods, including CNN-, ViT-, and CLIP-based methods. In general, CNN-based methods often tend to lower performance as shown in Table I and IV, showcasing the effectiveness of ViT and CLIP for a multi-modal fusion

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS IN MISSING MODALITY ON THE RGBNT201 DATASET. M(·) DENOTESTHE · MODAL ABSENCE

Methods	Structure	M(R)		M(N)		M(T)		M(RN)		M(RT)		M(NT)	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
HACNN [46]	CNN	12.5	11.1	20.5	19.4	16.7	13.3	9.2	6.2	6.3	2.2	14.8	12.0
MUDeep [49]	CNN	19.2	16.4	20.0	17.2	18.4	14.2	13.7	11.8	11.5	6.5	12.7	8.5
OSNet [7]	CNN	19.8	17.3	21.0	19.0	18.7	14.6	12.3	10.9	9.4	5.4	13.0	10.2
MFLN [47]	CNN	20.2	18.9	21.1	19.7	17.6	11.1	13.2	12.1	8.3	3.5	13.1	9.1
CAL [50]	CNN	21.4	22.1	24.2	23.6	18.0	12.4	18.6	20.1	10.0	5.9	17.2	13.2
PCB [51]	CNN	23.6	24.2	24.4	25.1	19.9	14.7	20.6	23.6	11.0	6.8	18.6	14.4
PFNet [12]	CNN	-	-	31.9	29.8	25.5	25.8	-	-	-	-	26.4	23.4
DENet [21]	CNN	-	-	35.4	36.8	33.0	35.4	-	-	-	-	32.4	29.2
TOP-ReID [23]	ViT	54.4	57.5	64.3	67.6	51.9	54.5	35.3	35.4	26.2	26.0	34.1	31.7
DeMo [13]	CLIP	63.3	65.3	72.6	75.7	56.2	54.1	**45.6**	**46.5**	26.3	24.9	40.3	38.5
MGRNet*	ViT	54.6	54.1	68.4	70.7	55.8	57.1	36.4	36.4	26.8	27.4	36.2	32.4
MGRNet†	CLIP	**66.3**	**68.8**	**75.2**	**78.1**	**58.6**	**59.0**	44.1	43.7	**30.7**	**29.8**	**42.6**	**42.0**

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS IN MISSING MODALITY ON THE RGBNT100 DATASET

Methods	Structure	M(R)		M(N)		M(T)		M(RN)		M(RT)		M(NT)	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
CCNet [22]	CNN	66.8	90.2	73.2	92.2	60.0	82.9	44.4	75.0	42.4	63.8	49.5	69.8
TOP-ReID [23]	ViT	70.6	90.6	77.9	94.5	64.0	81.5	42.5	69.3	45.9	65.4	55.4	77.8
DeMo [13]	CLIP	81.0	94.5	84.1	96.5	71.1	**87.6**	50.2	73.7	59.6	**78.1**	66.3	82.8
MGRNet*	ViT	75.2	93.3	79.6	94.9	66.7	83.4	42.9	67.1	50.3	70.7	59.5	78.9
MGRNet†	CLIP	**81.5**	**95.3**	**87.4**	**98.4**	**73.0**	87.1	**51.0**	**76.1**	**60.0**	75.6	**69.7**	**86.6**

strategy. For a fair and comprehensive comparison, we evaluate our proposed method using different vision encoders (e.g., ViT and CLIP) on four public benchmarks for multi-modal object ReID.

1) Comparison on RGBNT201 and Market1501-MM: As shown in Table I, one observation is that we achieve excellent performance compared to other methods for both ViT-based and CLIP-based MGRNet on two-person datasets. Notably, DeMo [13] achieves promising performance by considering the global information of each modality with the local information of all modalities. Meanwhile, IDEA [14] fuses semantic information to generate sampling local information, improving the identification ability. However, these methods treat all local patches equally and fail to distinguish patches of different quality. Compared with them, our approach can effectively pay attention to the quality differences of patches, extracting more discriminating features of persons. Eventually, for the RGBNT201 dataset [12], the ViT-based MGRNet achieves scores of 78.4%/82.5%, outperforming the sub-optimal ViT-based method by 4.7/2.0 percentage points (p.p.) for mAP/Rank-1, respectively. Additionally, although IDEA [14] introduces supplementary semantic information to enhance identification ability, our CLIP-based MGRNet consistently achieves higher results in mAP/Rank-1, demonstrating the robustness of its multi-modal representation learning. For the larger-scale generated Market1501-MM of multi-modal data [19], our method outperforms the sub-optimal results in mAP and Rank-1, which are higher than 2.4/0.5 p.p., and 1.6/0.5 p.p. for the ViT/CLIP-based MGRNet, respectively. The performance improvement is relatively modest compared with RGBNT201 [12], likely since it is a virtual multi-modal person ReID dataset generated by a GAN network [27]. Compared with real-world multi-modal datasets, synthetic data generally exhibit less realistic modality variations and a lower degree of challenge. This makes it difficult for the proposed MGRNet to fully prove its advantages in challenging multi-modal object ReID. As a result, although our method still achieves the best performance on Market1501-MM, the improvement is smaller than that on more challenging real-world datasets.

2) Comparison on RGBNT100 and MSVR310: To further validate the proposed method, we conduct additional experiments on vehicle ReID datasets. The results in Table IV demonstrate that MGRNet achieves superior performance with different vision encoders. In particular, the ViT-based MGRNet outperforms the recent ViT-based DeMo [13]. Furthermore, the CLIP-based variant consistently exceeds the semantic-guided CLIP-based IDEA [14], with improvements of 0.6/0.8 p.p. on RGBNT100 [17] and 6.2/4.8 p.p. on MSVR310 [22] in mAP/Rank-1, respectively. These results demonstrate the effectiveness of our method in addressing local feature quality inconsistency and alleviating the negative impact of low-quality local representations. They further confirm the ability of the proposed model to extract and integrate discriminative

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART METHODS ON VEHICLE REID DATASETS

	Methods	Structure	RGBNT100		MSVR310	
			mAP	R-1	mAP	R-1
Single	PCB [51]	CNN	57.2	83.5	23.2	42.9
	MGN [52]	CNN	58.1	83.1	26.2	44.3
	DMML [53]	CNN	58.5	82.0	19.1	31.1
	HRCN [54]	CNN	67.1	91.8	23.4	44.2
	AGW [1]	CNN	73.1	92.7	28.9	46.9
	OSNet [7]	CNN	75.0	95.6	28.7	44.8
	BoT [55]	CNN	78.0	95.1	23.5	38.4
	TransReID [9]	ViT	75.6	92.9	18.4	29.6
Multi	PFNet [12]	CNN	68.1	94.1	23.5	37.4
	IEEE [17]	CNN	61.3	87.8	21.0	41.0
	GAFNet [56]	CNN	74.4	93.4	-	-
	HAMNet [17]	CNN	74.5	93.3	27.1	42.3
	CCNet [22]	CNN	77.2	96.3	36.4	**55.2**
	GraFT [57]	ViT	76.6	94.3	-	-
	HTT [24]	ViT	75.7	92.6	-	-
	TOP-ReID [23]	ViT	81.2	96.4	35.9	44.6
	FACENet [58]	ViT	81.5	96.9	36.2	54.1
	EDITOR [25]	ViT	82.1	96.4	39.0	49.3
	RSCNet [26]	ViT	82.3	96.6	39.5	49.6
	DeMo [13]	ViT	82.4	96.0	39.1	48.6
	MGRNet*	ViT	**83.0**	**97.1**	**39.9**	49.3
	MambaPro [3]	CLIP	83.9	94.7	47.0	56.5
	DeMo [13]	CLIP	86.2	97.6	49.2	59.8
	IDEA [14]	CLIP	87.2	96.5	47.0	62.4
	MGRNet†	CLIP	**88.2**	**98.0**	**53.2**	**67.2**

cues for object ReID in complex and heterogeneous data environments.

D. Evaluation on Missing Modality

Table II and III evaluate the proposed GRMM strategy on RGBNT201 [12] and RGBNT100 [17], which present the experimental results in simulating missing modality scenarios. Our model consistently delivers strong performance. It is observed that our results outperform methods based on feature reconstruction, TOP-ReID [9], and DeMo [13]. For RGBNT201 [12], the average performance is higher by 2.0/2.2 p.p. and 0.9/2.9 p.p. in mAP and Rank-1 for ViT/CLIP-based methods. Meanwhile, we find that even in the absence of R/N modality, our results are still better than the single-modal methods as well as some multi-modal methods, especially EDITOR [25]. For RGBNT100 [17], the average performance improves by 9.33/8.07 p.p. in mAP and 5.68/5.13 p.p. in Rank-1 for ViT/CLIP-based methods, respectively. One observation is that even in the absence of N modality, our results are still more excellent than the compared methods, which are without modal absence. These indicate that graph reasoning can effectively reconstruct missing modal information by considering the essential structural relationships, yielding reliable multi-modal data in object ReID.

E. Ablation Study

To assess the effectiveness of the modules in our proposed MGRNet, we establish a baseline model as shown in Table V (a) that consists of the multi-branch vision encoders

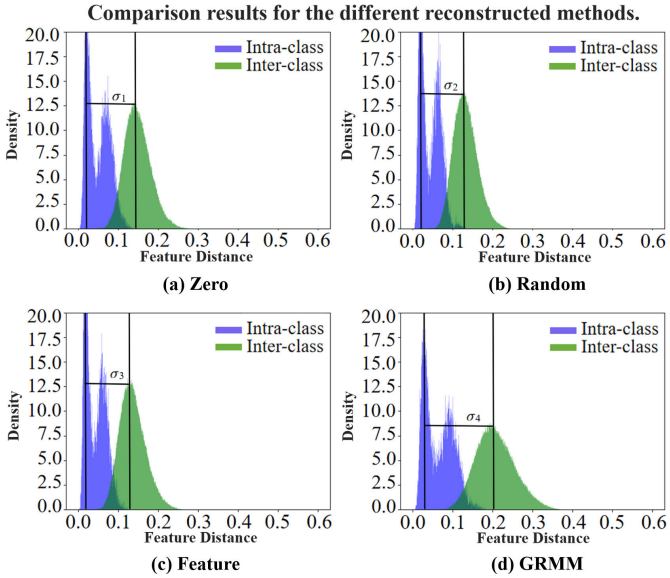

Fig. 4. The intra-class and inter-class distances of cross-modality features of different methods.

ViT/CLIP, Global-aware Multi-Head Attention, and the relevant optimization losses. We incrementally incorporate our proposed components into the baseline.

As shown in Table V (b), the integration of Modality-aware Graphs Learning (MGL) and Local-aware Graph Reasoning (LGR) enables the model to focus on critical local details for multi-modal data, obtaining a performance improvement in mAP and Rank-1, respectively. Furthermore, in Table V (c) and (d), the incorporation of the Selective Graph Nodes Swap (SGNS) significantly enhances performance, yielding a 2.7/5.2 p.p. and 2.8/5.7 p.p. improvement in mAP and Rank-1 compared to the ViT/CLIP-based baseline. This observation suggests that SGNS effectively mitigates the impact of multi-modal local noise. Lastly, as shown in Table V (e), (f), (h), our proposed Graph Reasoning on Missing Modality (GRMM) leverages both feature and structural relationships to restore modality-specific representations effectively. Here, $\mathcal{L}_F = \mathcal{L}_F^R + \mathcal{L}_F^N + \mathcal{L}_F^T$ and $\mathcal{L}_S = \mathcal{L}_S^R + \mathcal{L}_S^N + \mathcal{L}_S^T$ denote the feature-level and structural-level constraints, respectively, for multi-modal data reconstruction. The results in Table V (d), (e) and (i) indicate that both constraints help enhance the quality of missing-modality reconstruction and thus improve mAP and Rank-1 performance. Additionally, to evaluate the effectiveness of each step in the SGNS operation in Fig. 3, we further analyze the two selection operations. As can be seen in Table V, our proposed SGNS operation is effective. It can gradually alleviate the impact and noise caused by low-quality local features, thus further improving the model performance for multi-modal ReID tasks.

Summarily, integrating these components into the baseline model yields a notable improvement in experimental performance. The performance gains observed across various evaluation metrics validate the effectiveness of our proposed MGRNet. These results further emphasize the importance of structured graph-based fusion techniques in the field.

TABLE V
COMPARISON RESULTS FOR DIFFERENT MODULES ON THE RGBNT201 DATASET

	GRMI			GRMM		ViT				CLIP			
	MGL + LGR	SGNS^{1st}	SGNS^{2nd}	$\mathcal{L}_F$	$\mathcal{L}_S$	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
(a)	×	×	×	×	×	74.2	78.1	89.0	92.7	73.1	76.4	84.7	89.2
(b)	✓	×	×	×	×	75.0	78.5	88.8	93.1	74.6	79.1	86.7	90.9
(c)	✓	✓	×	×	×	75.7	79.5	90.8	94.0	75.8	80.3	87.9	91.0
(d)	✓	✓	✓	×	×	76.9	80.9	90.9	93.8	78.3	82.1	87.9	91.4
(e)	✓	✓	✓	✓	×	77.6	80.5	91.0	95.1	78.9	82.7	89.4	92.2
(f)	✓	×	×	✓	✓	77.2	80.5	88.9	93.5	75.2	79.3	89.2	91.6
(g)	✓	✓	×	✓	✓	78.0	81.0	88.5	92.6	79.9	83.3	89.5	92.1
(h)	✓	×	✓	✓	✓	78.2	80.3	89.1	94.3	76.1	81.1	89.0	91.7
(i)	✓	✓	✓	✓	✓	**78.4**	**82.5**	**90.9**	**95.2**	**80.5**	**85.0**	**90.0**	**92.6**

TABLE VI
COMPARISON RESULTS FOR DIFFERENT METHODS OF RECONSTRUCTION ON THE RGBNT201 DATASET

Modules	ALL		M(R)		M(N)		M(T)		M(RN)		M(RT)		M(NT)	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
Zero [28]	74.4	76.7	52.0	49.8	67.0	67.7	51.4	49.9	33.7	35.3	23.3	21.8	31.7	27.6
Random [28]	76.3	79.5	53.4	53.3	66.8	69.4	52.5	51.3	34.5	36.5	24.3	22.4	36.6	31.7
Feature [23]	76.4	78.7	54.2	52.2	66.1	69.0	52.4	53.0	37.3	**39.2**	25.7	27.3	34.4	30.0
GCRA [42]	74.9	77.6	48.5	47.1	64.7	64.8	**56.8**	56.8	35.9	37.0	26.4	**28.6**	**39.3**	**34.7**
AGDiC [43]	76.5	81.3	53.5	50.8	64.8	66.3	56.2	55.1	**37.4**	39.4	25.7	26.1	36.5	33.3
MGRNet*	**78.4**	**82.5**	**54.6**	**54.1**	**68.4**	**70.7**	55.8	**57.1**	36.4	36.4	**26.8**	27.4	36.2	32.4

F. Evaluation on GRMM

We conduct several experiments using reconstruction techniques of traditional methods [23], [28] and graph-based methods [42], [43] to validate the proposed GRMM strategy. By incorporating and modifying different reconstruction techniques, we evaluate the contribution of GRMM to the overall model performance under both complete-modality and missing-modality settings, as shown in Table VI. The results demonstrate that the proposed GRMM strategy consistently outperforms all comparison methods, confirming its crucial role in enhancing the model's robustness and accuracy. Meanwhile, the average performance across all configurations shows a notable improvement, highlighting the effectiveness and generalizability of our approach. Additionally, we visualize the inter-class and intra-class distances for different reconstruction methods as depicted in Fig. 4, where $\sigma_4 > \sigma_1, \sigma_2, \sigma_3$. This shows that the intra-class distance of GRMM is significantly reduced compared with other reconstruction methods.

G. Evaluation on Graph Reasoning

We conduct the comparative experiments with graph reasoning approaches [60], [61], [62], [63], including Graph Attention Network (GAT) [60], Graph Convolutional Network (GCN) [61], HORNet [62], HHGF [63]. These results are summarized in Fig. 5. The approaches first employ unshared multi-branch ViT to extract multi-modal features, and then construct a graph for all patches of each modality via Euclidean distance [38]. Finally, both label smoothing cross-entropy and triplet loss are combined to optimize the entire network. Experimental results show that MGRNet consistently surpasses traditional graph reasoning approaches [60], [61]. In addition, compared with recent graph reasoning methods [62], [63], MGRNet achieves further performance gains. This enhanced performance can be attributed to MGRNet's ability to alleviate the impact of low-quality local features, enhancing the discriminative information.

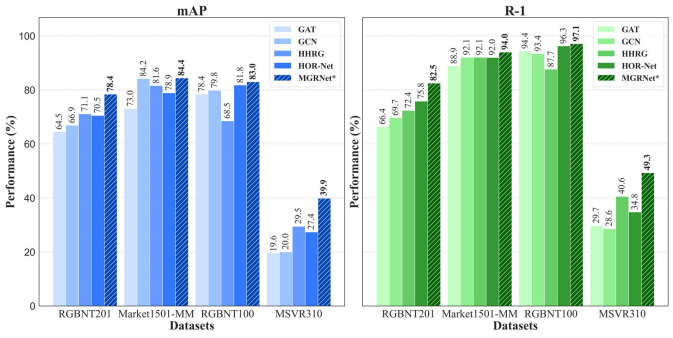

Fig. 5. Comparison of the graph reasoning methods on the common datasets.

H. Robustness Analysis

To comprehensively assess the model's robustness in complex real-world scenarios, we generate a noisy version of the RGBNT201 dataset by introducing arbitrary noise (e.g., occlusion, strong illumination, Gaussian noise, and salt-and-pepper noise) across multiple modalities. As shown in Table VII, these results demonstrate that our MGRNet consistently

TABLE VII

PERFORMANCE COMPARISON UNDER ARBITRARY NOISE AND LOW-RESOURCE ENVIRONMENTS ON THE RGBNT201 DATASET

Methods	Arbitrary Noise				Low Resource			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
TOP-ReID [23]	56.8	57.2	75.2	82.4	56.0	57.9	70.2	76.4
EDITOR [25]	57.4	60.8	77.2	84.1	47.2	45.8	62.9	72.9
DeMo [13]	70.5	72.4	85.6	89.8	60.8	64.5	**79.2**	**84.3**
MambaPro [3]	68.0	72.5	85.0	90.2	58.9	62.7	75.2	82.7
MGRNet*	65.5	69.4	80.9	86.0	58.7	59.1	73.1	78.7
MGRNet†	**72.1**	**76.4**	**88.2**	**91.9**	**62.5**	**64.6**	76.6	82.7

TABLE VIII

COMPARISON RESULTS FOR THE NUMBER OF SWAP-NODES ON THE RGBNT201 DATASET

	ViT				CLIP			
k	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
0	77.2	80.5	88.9	93.5	75.2	79.3	89.2	91.6
10	78.3	82.2	90.9	94.7	79.9	82.8	90.2	92.6
20	**78.4**	**82.5**	**90.9**	**95.2**	**80.5**	**85.0**	90.0	92.6
40	78.0	80.3	89.5	95.2	78.6	82.5	**90.7**	**92.9**
60	77.0	79.4	90.1	94.9	76.8	81.0	87.3	89.7
80	61.6	60.3	76.2	86.0	76.6	80.0	87.8	91.4

achieves superior performance compared with state-of-the-art approaches, highlighting its robustness. Furthermore, we simulate extremely low-resource environments by reducing the input image resolution from $256 \times 128 \times 3$ to $128 \times 64 \times 3$. As shown in Table VII, our method maintains a competitive performance, confirming its effectiveness in resource-constrained settings.

I. Hyperparameter Analysis

We further analyze the influence of the swap-node number (hyperparameter k) on the performance of our model. The results are displayed in Table VIII, where the first row represents the case where the SGNS method is not applied in different vision encoders. From the results, we observe a trend: as the k-value increases, the model's performance initially improves before reaching a peak and then declines. When $k = 0$, low-quality nodes remain uncorrected. Increasing k improves performance by replacing more unreliable nodes with complementary cues, whereas an overly large k obscures original modality information and degrades performance. This behavior highlights that an optimal number of swapped nodes enhances the model's ability to capture rich local features and reduce the impact of low-quality local features. These results prove that our SGNS operation is more effective than using a direct graph-based approach.

J. Visualization

1) Discriminative Attention Maps: We present visualization via gradually adding our proposed strategies, utilizing Grad-CAM [64], which demonstrates the ability of our MGRNet to effectively capture the relevant regions of the input images. Compared to the proposed MGRNet without the GRMI and GRMM strategies, our method can capture more critical areas

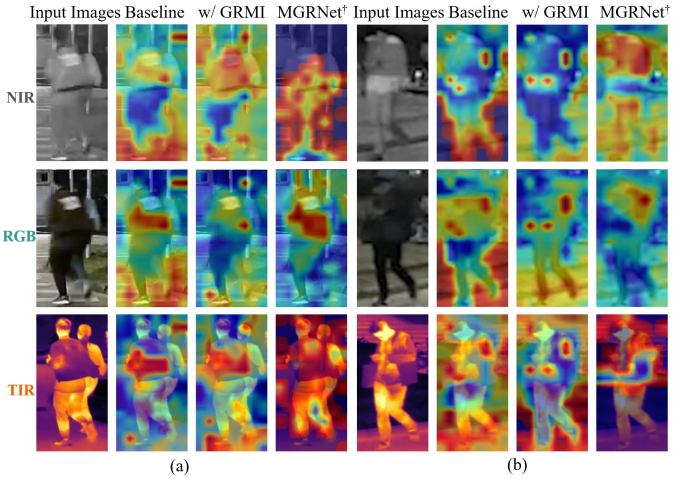

Fig. 6. Visualization results using Grad-CAM of models with progressively incorporated GRMI and GRMM.

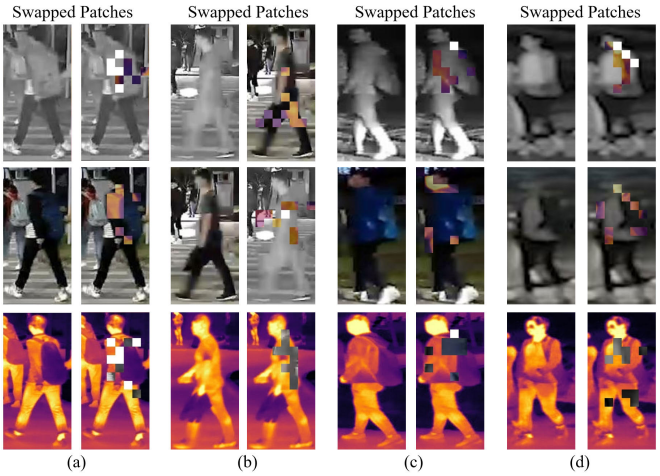

Fig. 7. Visualization results of swapped patches of Graph Reasoning on Modal Interaction (GRMI) on the RGBNT201 dataset.

while significantly reducing the impact of irrelevant or noisy regions. The results in Fig. 6 (a) and (b) show that our method can capture key information more comprehensively, while this method has a better effect in background noise suppression, thus improving the quality of overall feature representation.

Additionally, to further evaluate the GRMI strategy, we visualize the exchanged patches via the SGNS operation as shown in Fig. 7. By exchanging local features, problems related to low-quality features, such as the face and bag, are solved, enhancing the representation of discriminative information and improving the effectiveness of modal representation.

2) Reconstruction Feature: To more intuitively assess the GRMM strategy, we utilize feature maps to visualize different modal images. As shown in Fig. 8, our model can recover superior feature representations in cases where modalities are missing. It is known that RGB contains more color and detail information, covering the brightness and edge features of NIR, while NIR is more prominent in low light structure and contrast, and is a complement to RGB [65]. It is found that the associated region in the NIR modality when generating the RGB modality is less accurate than the real one, as

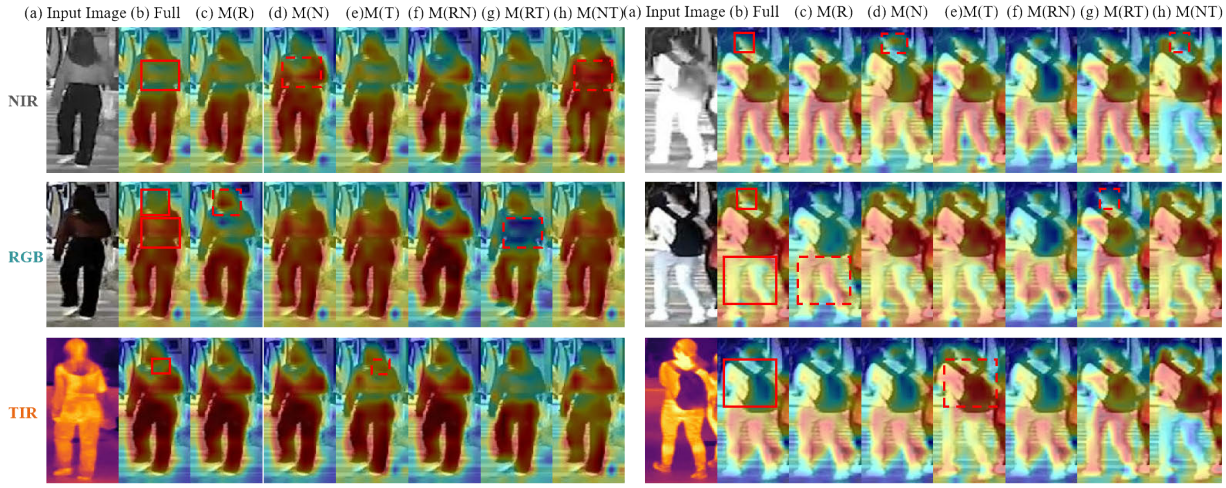

Fig. 8. Feature map for complete and missing modality on the RGBNT201 dataset.

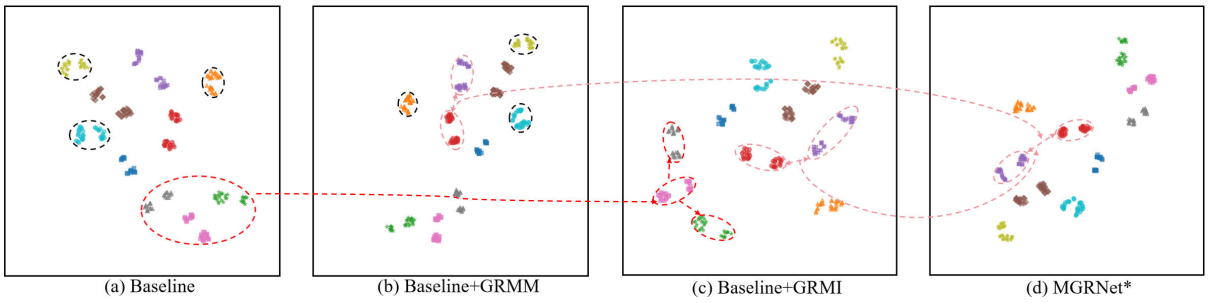

Fig. 9. Feature distribution on different strategies by using t-SNE on RGBNT201. The different colors represent different identities.

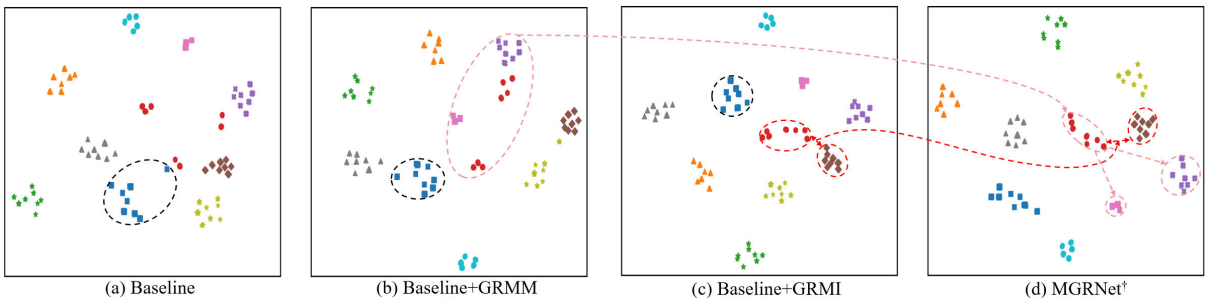

Fig. 10. Feature distribution on different strategies by using t-SNE on RGBNT100. The different colors represent different identities.

shown in Fig. 8 (g), while the associated region in the RGB modality when generating the NIR modality is more accurate than the real one, as shown in Fig. 8 (h). In the future, we will consider the specific information and complementary information between modalities to further optimize our work.

3) Feature Distribution: We visualize the obtained representations by the t-SNE tool [66]. Fig. 9/10 (a) is the visualized result by leveraging the ViT/CLIP-based baseline for RGBNT201/RGBNT100. Fig. 9/10 (b) and (c) add our proposed GRMM and GRMI strategies into this baseline, respectively. One can observe that GRMM can reduce the distance between intra-class by minimizing the disparity between modalities, and GRMI can increase the distance between inter-class by reducing the impact of low-quality patches. Moreover, in Fig. 9/10 (d), it can be found that our proposed MGRNet exhibits larger inter-class distances and smaller intra-class distances than the baseline, indicating that our model is better at capturing differences between objects.

4) Failure Cases Analysis: To further evaluate the retrieval performance of MGRNet in real-world scenarios, we visualize several failure cases on the RGBNT201 dataset [12]. Benefiting from its feature interaction and reconstruction capability, MGRNet achieves significant performance improvements. However, as shown in Fig. 11, MGRNet still faces difficulties in scenarios with extreme lighting degradation, background occlusion, and modal noise, which hinder its ability to focus on object semantics. In the future, we plan to investigate more robust and fine-grained multi-modal graph reasoning approaches to overcome these limitations.

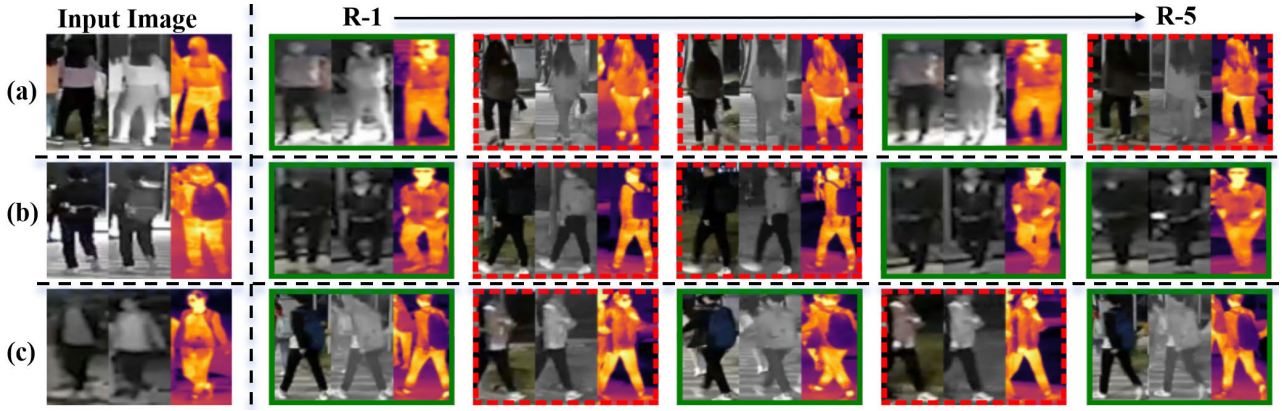

Fig. 11. Visualization results of top-5 failure cases on the RGBNT201 dataset.

V. Conclusion

In this paper, we propose a novel approach, named Modality-aware Graph Reasoning Network (MGRNet), which effectively enhances information interactions and recovers missing modalities for multi-modal object ReID. First, we introduce the construction of modality-aware graphs to adaptively model the relationships among local features, learning important local details. Second, the selective graph nodes swap operation is introduced to alleviate the impact of low-quality features from the modalities while effectively capturing crucial local details, promoting multi-modal fusion. Finally, we feed the swapped modality-aware graphs into the local-aware graph reasoning module to achieve message propagation, thus yielding a reliable feature representation of multi-modal data in object ReID. Additionally, we propose that the MGRNet is capable of recovering missing modalities based on their structural relationships, effectively reducing and minimizing multi-modal disparity. Overall, the proposed MGRNet achieves state-of-the-art performance on multi-modal ReID datasets. MGRNet serves as a preliminary framework for multi-modal fusion and recovery in graph reasoning, demonstrating promising results. However, multi-modal fusion remains a challenging task. In the future, our work will focus on developing more efficient and fine-grained fusion strategies of graph reasoning.

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[2] R. Sun, L. Chen, L. Zhang, R. Xie, and J. Gao, "Robust visible-infrared person re-identification based on polymorphic mask and wavelet graph convolutional network," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2800–2813, 2024.

[3] Y. Wang et al., "Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 8, 2025, pp. 8150–8158.

[4] H. Rao et al., "Self-supervised gait encoding with locality-aware attention for person re-identification," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 898–905.

[5] Z. Lu, R. Lin, and H. Hu, "Modality and camera factors bi-disentanglement for NIR-VIS object re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1989–2004, 2023.

[6] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 677–683.

[7] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3701–3711.

[8] Y. Hao, N. Wang, X. Gao, J. Li, and X. Wang, "Dual-alignment feature embedding for cross-modality person re-identification," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 57–65.

[9] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15013–15022.

[10] P. Wang et al., "DRFormer: A discriminable and reliable feature transformer for person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 980–995, 2025.

[11] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 4610–4617.

[12] A. Zheng, Z. Wang, Z. Chen, C. Li, and J. Tang, "Robust multi-modality person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, vol. 35, no. 4, pp. 3529–3537.

[13] Y. Wang, Y. Liu, A. Zheng, and P. Zhang, "Decoupled feature-based mixture of experts for multi-modal object re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 8, 2025, pp. 8141–8149.

[14] Y. Wang, Y. Lv, P. Zhang, and H. Lu, "IDEA: Inverted text with cooperative deformable aggregation for multi-modal object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 29701–29710.

[15] Z. Wei, X. Yang, N. Wang, and X. Gao, "Dual-adversarial representation disentanglement for visible infrared person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2186–2200, 2024.

[16] H. Rao et al., "A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6649–6666, Oct. 2022.

[17] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, "Multi-spectral vehicle re-identification: A challenge," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, vol. 34, no. 7, pp. 11345–11353.

[18] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 229–247.

[19] Z. Wang, C. Li, A. Zheng, R. He, and J. Tang, "Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 2633–2641.

[20] J. Crawford, H. Yin, L. McDermott, and D. Cummings, "UniCat: Crafting a stronger fusion baseline for multimodal re-identification," 2023, *arXiv:2310.18812*.

[21] A. Zheng, Z. He, Z. Wang, C. Li, and J. Tang, "Dynamic enhancement network for partial multi-modality person re-identification," 2023, *arXiv:2305.15762*.

[22] A. Zheng, X. Zhu, Z. Ma, C. Li, J. Tang, and J. Ma, "Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark," *Inf. Fusion*, vol. 100, Dec. 2023, Art. no. 101901.

[23] Y. Wang, X. Liu, P. Zhang, H. Lu, Z. Tu, and H. Lu, "TOP-ReID: Multi-spectral object re-identification with token permutation," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 5758–5766.

[24] Z. Wang, H. Huang, A. Zheng, and R. He, "Heterogeneous test-time training for multi-modal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 5850–5858.
[25] P. Zhang, Y. Wang, Y. Liu, Z. Tu, and H. Lu, "Magic tokens: Select diverse tokens for multi-modal object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17117–17126.
[26] Z. Yu et al., "Representation selective coupling via token sparsification for multi-spectral object re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3633–3648, Apr. 2025.
[27] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.
[28] C. Wang et al., "Cross-modal pattern-propagation for RGB-T tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7062–7071.
[29] H. Liu, D. Xia, and W. Jiang, "Towards homogeneous modality learning and multi-granularity information exploration for visible-infrared person re-identification," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 3, pp. 545–559, May 2023.
[30] X. Liu, W. Liu, J. Zheng, C. Yan, and T. Mei, "Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 907–915.
[31] F. Shen, X. Peng, L. Wang, X. Hao, M. Shu, and Y. Wang, "HSGM: A hierarchical similarity graph module for object re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2022, pp. 1–6.
[32] B. X. Nguyen, B. D. Nguyen, T. Do, E. Tjiputra, Q. D. Tran, and A. Nguyen, "Graph-based person signature for person re-identifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3487–3496.
[33] H. Liu, Z. Xiao, B. Fan, H. Zeng, Y. Zhang, and G. Jiang, "PrGCN: Probability prediction with graph convolutional network for person re-identification," *Neurocomputing*, vol. 423, pp. 57–70, Jan. 2021.
[34] D. Cheng, H. Tai, N. Wang, C. Fang, and X. Gao, "Neighbor consistency and global-local interaction: A novel pseudo-label refinement approach for unsupervised person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 9070–9084, 2024.
[35] B. Jiang, X. Wang, A. Zheng, J. Tang, and B. Luo, "PH-GCN: Person retrieval with part-based hierarchical graph convolutional network," *IEEE Trans. Multimedia*, vol. 24, pp. 3218–3228, 2022.
[36] Z. He, H. Zhao, and W. Feng, "PGGANet: Pose guided graph attention network for person re-identification," 2021, *arXiv:2111.14411*.
[37] Y. Lv, G. Wang, W. Zhao, W. Zhao, and Z. Guan, "Edge-weight-embedding graph convolutional network for person reidentification," *IEEE Intell. Syst.*, vol. 39, no. 4, pp. 74–82, Jul. 2024.
[38] Z. Liu, H. Li, R. Li, Y. Zeng, and J. Ma, "Graph embedding based on Euclidean distance matrix and its applications," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 1140–1149.
[39] Y. Ye and S. Ji, "Sparse graph attention networks," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 905–916, Jan. 2023.
[40] C. Zheng et al., "Robust graph representation learning via neural sparsification," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11458–11468.
[41] S. Miao, M. Liu, and P. Li, "Interpretable and generalizable graph learning via stochastic attention mechanism," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 15524–15543.
[42] G. Du, T. Gong, and L. Zhang, "Graph-based consistent reconstruction and alignment for imbalanced text–image person re-identification," *Expert Syst. Appl.*, vol. 260, Jan. 2025, Art. no. 125429.
[43] Y. Li, Y. Dong, Y. Wu, H. Yan, and L. Gao, "Alzheimer's disease recognition based on adaptive graph normalization flow for incomplete multimodal data fusion," in *Proc. Med. Image Comput. Comput. Assist. Intervent*, 2025, pp. 64–73.
[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[45] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.
[46] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.
[47] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2109–2118.
[48] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[49] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5409–5418.
[50] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1005–1014.
[51] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *Proc. Eur. Conf. Comput. Vis.*, 2017, pp. 501–518.
[52] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.
[53] G. Chen, T. Zhang, J. Lu, and J. Zhou, "Deep meta metric learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9546–9555.
[54] J. Zhao, Y. Zhao, J. Li, K. Yan, and Y. Tian, "Heterogeneous relational complement for vehicle re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 205–214.
[55] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1487–1495.
[56] J. Guo, X. Zhang, Z. Liu, and Y. Wang, "Generative and attentive fusion for multi-spectral vehicle re-identification," in *Proc. 7th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Apr. 2022, pp. 1565–1572.
[57] H. Yin, J. Li, E. Schiller, L. McDermott, and D. Cummings, "GraFT: Gradual fusion transformer for multimodal re-identification," 2023, *arXiv:2310.16856*.
[58] A. Zheng, Z. Ma, Y. Sun, Z. Wang, C. Li, and J. Tang, "Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification," *Inf. Fusion*, vol. 116, Apr. 2025, Art. no. 102800.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–20.
[60] P. Veličković et al., "Graph attention networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.
[61] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.
[62] L. Qiu, S. Chen, Y. Yan, J.-H. Xue, D.-H. Wang, and S. Zhu, "High-order structure based middle-feature learning for visible-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4596–4604.
[63] Y. Feng et al., "Homogeneous and heterogeneous relational graph for visible-infrared person re-identification," *Pattern Recognit.*, vol. 158, Feb. 2025, Art. no. 110981.
[64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[65] L. Yan, X. Wang, M. Zhao, S. Liu, and J. Chen, "A multi-model fusion framework for NIR-to-RGB translation," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2020, pp. 459–462.
[66] D. Kobak and G. C. Linderman, "Initialization is critical for preserving global data structure in both t-SNE and UMAP," *Nature Biotechnol.*, vol. 39, no. 2, pp. 156–157, Feb. 2021.

Xixi Wan received the B.A. degree from Wannan Medical University, Wuhu, China, in 2022. She is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Anhui University, Hefei, China. Her current research interests are in computer vision and machine learning, especially multi-modal learning.

Aihua Zheng received the B.Eng. and joint master's and Ph.D. degrees in computer science and technology from Anhui University, China, in 2006 and 2008, respectively, and the Ph.D. degree in computer science from the University of Greenwich, U.K., in 2012. She visited the University of Stirling and Texas State University in 2013 and 2019, respectively. She is currently a Full Professor and the Ph.D. Supervisor with the School of Artificial Intelligence, Anhui University. Her main research interests include vision-based artificial intelligence and pattern recognition, especially on object re-identification, audio visual computing, and multi-modal intelligence.

Zi Wang received the B.Eng. and joint master's and Ph.D. degrees from the School of Computer Science and Technology, Anhui University. He is currently with the School of Biomedical Engineering, Anhui Medical University. He is primarily engaged in research on computer vision, medical image processing, and multi-modal learning.

Bo Jiang received the B.S. degree in mathematics and applied mathematics and the M.Eng. and Ph.D. degrees in computer science from Anhui University, China, in 2009, 2012, and 2015, respectively. He is currently an Associate Professor of computer science with Anhui University. He has published more than 100 papers, including 50 papers in top conferences and journals, such as CVPR, NeurIPS, IJCV, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. His current research interests include image matching, graph data representation, and learning.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor and the Ph.D. Supervisor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, pattern recognition, and machine learning.

Jixin Ma received the B.Sc. and M.Sc. degrees in mathematics in 1982 and 1988, respectively, and the Ph.D. degree in computer sciences in 1994. He is currently a Full Professor and the Director of the Ph.D./M.Phil. Program, School of Computing and Mathematical Sciences, University of Greenwich, U.K. His research interests include temporal logic, temporal databases, reasoning about action and change, case-based reasoning, pattern recognition, machine learning, and information security.