HOH-Net: High-Order Hierarchical Middle-Feature Learning Network for Visible-Infrared Person Re-Identification

Liuxiang Qiu, Si Chen, *Senior Member, IEEE*, Jing-Hao Xue, *Senior Member, IEEE*, Da-Han Wang, *Member, IEEE*, Shunzhi Zhu, Yan Yan, *Senior Member, IEEE*

***Abstract*—Visible-infrared person re-identification (VI-ReID) is a cross-modality retrieval task that aims to match images of the same person across visible (VIS) and infrared (IR) modalities. Existing VI-ReID methods ignore high-order structure information of features and struggle to learn a reliable common feature space due to the modality discrepancy between VIS and IR images. To alleviate the above issues, we propose a novel high-order hierarchical middle-feature learning network (HOH-Net) for VI-ReID. We introduce a high-order structure learning (HSL) module to explore the high-order relationships of short- and long-range feature nodes, for significantly mitigating model collapse and effectively obtaining discriminative features. We further develop a fine-coarse graph attention alignment (FCGA) module, which efficiently aligns multi-modality feature nodes from node-level and region-level perspectives, ensuring reliable middle-feature representations. Moreover, we exploit a hierarchical middle-feature agent learning (HMAL) loss to hierarchically reduce the modality discrepancy at each stage of the network by using the agents of middle features. The proposed HMAL loss also exchanges detailed and semantic information between low- and high-stage networks. Finally, we introduce a modality-range identity-center contrastive (MRIC) loss to minimize the distances between VIS, IR, and middle features. Extensive experiments demonstrate that the proposed HOH-Net yields state-of-the-art performance on the image-based and video-based VI-ReID datasets. The code is available at: https://github.com/Jaulaucoeng/HOS-Net.**

***Index Terms*—Visible-infrared person re-identification, high-order structure, middle-feature learning.**

I. Introduction

PERSON re-identification (ReID) [1]–[3] has drawn more and more attention in recent years because of its critical role in security and surveillance. Visible-infrared person re-identification (VI-ReID) leverages both visible (VIS) and infrared (IR) cameras to match pedestrian images across the bright and low-light conditions. The VI-ReID methods not only mitigate the problems of single-modality ReID (e.g., occlusion and posture deformation), but also need to handle the modality discrepancy between VIS and IR images.

To bridge the modality gap, existing VI-ReID methods can be classified as image-level and feature-level. Image-level methods [4], [5] often employ generative adversarial networks (GANs [6]) to generate middle or new modality images. For instance, to mitigate the modality discrepancy, Wei et al. [4] introduced a reciprocal bidirectional framework that generates the middle modality images from the latent space by translating two opposite mappings between VIS and IR modalities by the generative adversarial network. However, GAN-based methods easily encounter issues such as color inconsistency or the loss of image details, which make the generated images less reliable for training and subsequent retrieval.

Feature-level methods [1], [7]–[10] typically adopt a two-step learning process. First, these methods extract VIS and IR feature maps using weight-specific sub-networks separately. Subsequently, the weight-shared feature extraction projects these modality-specific features into a common feature space. For instance, Liang et al. [11] developed a pure Transformer network to capture long-range information from different modalities with the modality-aware enhancement loss. To enhance feature representation, Zhang et al. [9] have attempted to introduce the self-distillation to consistently focus on discriminative regions from the high-stage to the low-stage for modality feature learning. The above feature-level methods generally have three shortcomings. First, they often neglect the high-order structural information of features, such as the complex dependencies across feature nodes, which are crucial for retrieving cross-modality images. Second, the traditional methods extract person features from the low-stage to the high-stage or enhance the feature representation by distillation from the high-stage to the low-stage. Such strategies ignore the bi-directional interaction between the low-stage and the high-stage and are thus hard to explore the detailed and semantic features. Third, existing methods try to directly minimize the distances between VIS and IR features, or generate the auxiliary features from one or two modalities to mitigate the modality discrepancy, but they still lack efficient alignment

This work was supported in part by the National Natural Science Foundation of China (No. 62372388); the Natural Science Foundation of Xiamen (No. 3502Z202573073); the Unveiling and Leading Projects of Xiamen (No. 3502Z20241011); the Major Science and Technology Plan Project on the Future Industry Fields of Xiamen City (No. 3502Z20241027); the Open Project of the State Key Laboratory of Multimodal Artificial Intelligence Systems (No. MAIS2024101). (*Corresponding author: Si Chen.*)

Liuxiang Qiu, Si Chen, Da-Han Wang, and Shunzhi Zhu are with the Fujian Key Laboratory of Pattern Recognition and Image Understanding, School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China, and Liuxiang Qiu is also with the School of Informatics, Xiamen University, Xiamen 361005, China (email: liuxiangqiu007@gmail.com; chensi@xmut.edu.cn; wangdh@xmut.edu.cn; szzhu@xmut.edu.cn).

Jing-Hao Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK (e-mail: jinghao.xue@ucl.ac.uk).

Yan Yan is with the School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: yanyan@xmu.edu.cn).

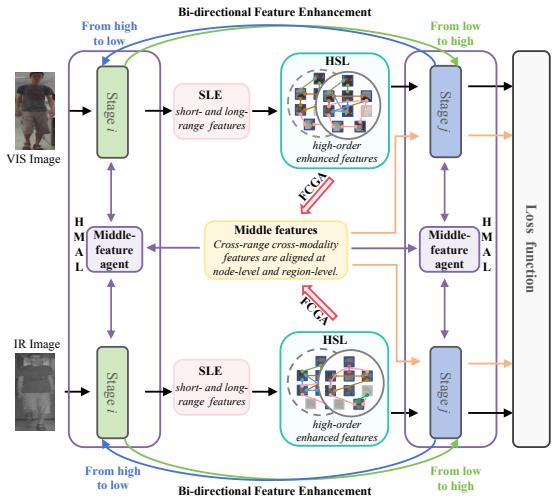


Fig. 1. Illustration of the proposed method. Our HOH-Net aligns different modalities and ranges of high-order enhanced features at node-level and region-level simultaneously to generate the reliable middle-feature agents, and leverages the bi-directional feature enhancement to hierarchically reduce the modality discrepancy.

and full utilization between different modality features.

To address the above issues, we propose a novel high-order hierarchical middle-feature learning network (HOH-Net), which is shown in Fig. 1. The HOH-Net is made up of a high-order structure learning (HSL) module, a fine-coarse graph attention alignment (FCGA) module, a hierarchical middle-feature agent learning (HMAL) loss, and a modality-range identity-center contrastive (MRIC) loss for VI-ReID. The key innovation of our method lies in the novel formulation of exploiting high-order structure information and hierarchical middle-feature learning to learn a discriminative and reliable common feature space, thereby significantly mitigating the modality gap.

Specifically, given a VIS-IR image pair, the HSL module captures the high-order relationships between the short-range and long-range features that are extracted from the short- and long-range feature extraction (SLE) module using a whitened hypergraph. Instead of directly adding or concatenating features from different modalities and ranges, we design an FCGA module that aligns these features appropriately and effectively at node-level and region-level simultaneously to achieve reliable middle features. Besides, we propose a HMAL loss to address the modality gap hierarchically by utilizing middle-feature agents and executing bi-directional interactions between different stages to enhance feature representation. Finally, we reduce the distances among VIS, IR, and middle center features by an MRIC loss, thereby smoothing the learning process of the common feature space between modalities. On the SYSU-MM01, RegDB, LLCM, and HITSZ-VCM datasets, our method achieves impressive 76.2%, 95.1%, 65.7%, and 74.8% in Rank-1, respectively.

The main contributions of our work are as follows:

- We propose an HSL module to learn high-order structure information of both short and long-range features. Such a novel way effectively models high-order relationships across different feature nodes of each pedestrian image and avoids the problem of model collapse.
- We design a lightweight yet effective FCGA module that can refine the details of each high-order node-level feature and perceive the semantic association of region-level features simultaneously to achieve reliable middle features.
- An HMAL loss is designed to hierarchically reduce the modality discrepancy at each stage network by the middle-feature agents and perform the bi-directional feature enhancement between different stages to enhance the detailed representation and the semantic relationship of features.
- An MRIC loss is designed to minimize the distances between VIS, IR, and middle features in the embedding space. This is beneficial to extracting discriminative modality-shared pedestrian features.

This paper significantly extends our previous conference work HOS-Net [12]. The limitations of our previous work include the following: First, the computational cost of generating the middle features through graph attention is high and did not make full use of the middle features. Second, the previous method extracted modality-shared features from the low stage to the high stage, ignoring the importance of bi-directional interaction between different stages that can enhance feature representation. The HOH-Net addresses these limitations in two main ways. (1) We further develop a fine-coarse graph attention alignment (FCGA) module to refine the high-order node-level features and perceive the contextual relationship between region-level features to achieve more reliable middle features with less model complexity. (2) We design an HMAL loss to mitigate modality discrepancy from a hierarchical view by introducing the agents of the middle features at each VIS and IR modality-shared feature extraction stage. The proposed HMAL loss also enables the bi-directional interaction of features between different stages, for obtaining richer semantic and more detailed feature information than the previous HOS-Net. In the experiments, we also provide more comprehensive experimental evaluations, including comparative experiments, ablation studies, parameter analyses, and visualization analyses. Compared to the previous HOS-Net, the HOH-Net achieves lower computational cost and superior retrieval accuracy than our previous work (the number of parameters of the HOH-Net is reduced by 29.5%) and the Rank-1 of our method is improved by 0.6%, 0.4%, and 0.8% on the three image-based VI-ReID datasets, i.e., SYSU-MM01, RegDB, and LLCM, respectively. In addition, our method can also be easily extended to the video-based VI-ReID field, and compared to the existing video-based methods, our HOH-Net achieves the best 74.8% Rank-1 on the HITSZ-VCM dataset.

II. Related Work

A. Visible-Infrared Person Re-Identification (VI-ReID)

VI-ReID methods can be divided into image-level and feature-level methods to reduce the modality discrepancy. The

image-level methods [4], [5], [13] often minimize the modality gap by generating middle-modality images or new modality images. Wang et al. [13] attempted to introduce a generative adversarial network to generate new modality images from VIS and IR modalities by jointly aligning the pixel-level and feature-level features. Liu et al. [5] proposed a two-stage modality enhancement network to perform the cross-modality style translation and optimized the structures of images for VI-ReID. Besides, Li et al. [14] leveraged the anaglyph data of the pedestrian as the middle modality images to reduce the modality gap. Du et al. [15] proposed a channel-blended transformation mechanism to confuse the VIS and IR information and reduce the influence of modality-specific features, thereby facilitating the learning of modality-shared features. However, image-level methods easily encounter issues such as color inconsistency or the loss of image details when generating images by the generative adversarial network (GAN), which is less reliable for training and subsequent visible-infrared retrieval.

The feature-level methods seek to reduce the modality discrepancy by mapping the features of different modalities into a common feature space. A few methods [1], [8], [16] leverage the weight-shared CNN or ViT as the backbone to extract modality-shared features. Hybrid models of CNN and Transformer [10], [17]–[20] can effectively extract short-range and long-range features. For example, Zhao et al. [10] enhanced the spatial-channel information of the pedestrian by adopting the CNN-Transformer hybrid network. Chen et al. [20] attempt to introduce the off-the-shelf key point extractors (e.g., OpenPose [21]) to generate key point labels of person images and achieve features based on the CNN-Transformer hybrid network, aiming to learn modality-irrelevant features. But the key point extractor may bring noisy labels, deteriorating the discriminability of final ReID features. However, the above feature-level methods neglect the high-order structure information of features (i.e., the complex and diverse relationships across features) that is important for VI-ReID. To solve the above problem, our work introduces the high-order structure learning to obtain the high-order relationships between the short- and long-range features and avoid the model collapse by a whitened hypergraph.

To obtain a common feature space, a lot of feature-level methods [1], [8], [22]–[24] employ the contrastive-based loss that directly minimizes the distances between VIS and IR features. However, it is not a trivial task to learn a reliable common feature space due to the large modality discrepancy between modalities. Different from these methods that tend to minimize the distances between VIS and IR features directly, Zhang et al. [25] tried to generate diverse VIS or IR embeddings for learning informative feature representations to mitigate the modality gap. Jiang et al. [26] adopted the modality-level and instance-level alignments for learning robust modality compensation. Li et al. [27] introduced the cross-modality semantic alignment to explore the inter-modality correlation for eliminating the modality discrepancy. However, they ignored the importance of fine-coarse alignment for generating reliable middle features from different modalities and ranges to narrow the difference between VIS and IR images. Different from these methods, our method generates reliable hierarchical middle-feature agents via the fine-coarse graph attention alignment, greatly promoting our method to learn a discriminative and reliable common feature space.

In addition, to improve the discriminative ability of the network, Yang et al. [28] designed a saliency response module that adopts the location attention mechanism to build contextual connections between person features. Tian et al. [29] adopted the variational self-distillation to fit the mutual information between the input feature and its representation, thus obtaining the multi-view information for VI-ReID. The above methods follow low-to-high feature extraction, which ignores the interaction between features at different stages. To this end, the proposed HOH-Net performs the bi-directional enhancement between different stages to enhance the detailed representation and the semantic relationship of features. Moreover, we reduce the distances among VIS, IR, and middle center features by a modality-range identity-center contrastive loss, thereby smoothing the learning process of the common feature space between ranges and modalities.

B. Graph Neural Network

Graph neural network (GNN) is a type of neural network to process graph-structured data. Zhang et al. [30] adopted the GNN to select correlated nodes for information aggregation, thereby establishing the robust connection between the target and the search regions. Zhang et al. [31] introduced the GNN to perform the progressive relationship-mining for text-to-image ReID. Contrasting with the vanilla graph models that only allow connections between two nodes, Feng et al. [32] proposed the novel hypergraph neural network (HGNN) to represent high-order feature correlations by utilizing a hypergraph structure. Wadhwa et al. [33] adopted the HGNN to learn the complex relationship among the incomplete features for the image inpainting. Han et al. [34] utilized the power of the hypergraph to encode image information and update the hypergraph structure by the fuzzy c-means method that can reduce the computational burden. Nevertheless, the above methods that rely on the HGNN may easily suffer from the model collapse (i.e., complex and diverse high-order correlations collapse to a single correlation) since the small differences in the feature nodes of pedestrians and the hyperedge can connect an arbitrary number of nodes. Different from the above methods, this paper introduces the whitening operation to HGNN, which can play the role of "scattering" on the nodes of the hypergraph, thereby significantly alleviating model collapse.

Besides, to establish the correspondence between feature nodes, several methods [35]–[37] attempt to introduce the graph attention network (GAT) to enhance the representation of features. For instance, Dong et al. [35] fused the characteristics of CNN and GAT to discover feature connections for hyperspectral image classification. However, the above methods consider the correspondence between feature nodes at node-level, and ignore the semantic connections between region-level features that can encapsulate the context of features. In this work, we develop a fine-coarse graph attention alignment

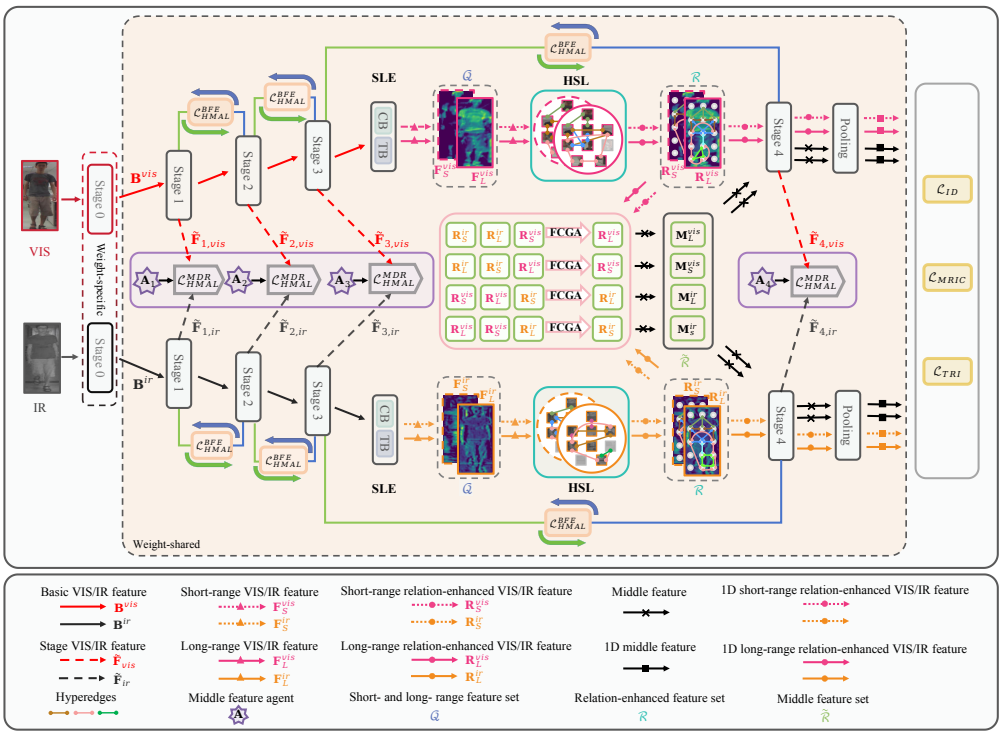


Fig. 2. Overall of the proposed HOH-Net, including a high-order structure learning (HSL) module and a fine-coarse graph attention alignment (FCGA) module. The HOH-Net is jointly optimized by $\mathcal{L}_{CE}$, $\mathcal{L}_{TRI}$, and a hierarchical middle-feature agent learning (HMAL) loss $\mathcal{L}_{HMAL}$, and a modality-range identity-center contrastive loss $\mathcal{L}_{MRIC}$.

(FCGA) module to leverage high-order node-level and region-level features for achieving reliable middle features. Besides, we also generate the middle-feature agents to hierarchically mitigate the modality gap at each modality-shared feature extraction stage by introducing a hierarchical middle-feature agent learning (HMAL) loss.

III. Proposed Method

A. Overview

The overall of our proposed HOH-Net is given in Fig. 2. The HOH-Net mainly consists of an HSL module and an FCGA module with the HMAL loss and the MRIC loss. In this paper, we adopt a two-stream AGW [1] as the backbone. Firstly, we feed the VIS-IR image pair with the same identity to the backbone for obtaining paired VIS-IR features. Then, the HSL module introduces a whitened hypergraph network to exploit high-order structure information of short-range and long-range features that are obtained from the short- and long-range feature extraction (SLE) module. Furthermore, the FCGA module aligns different modalities and ranges of features to generate reliable middle features effectively at the node-level and region-level. The HMAL loss mitigates the modality discrepancy hierarchically based on the middle-feature agents and can constrain the bi-directional interaction between different stages to improve the representation of features. Besides, in the embedding space, we develop an MRIC loss to reduce the distances between the VIS, IR, and middle features, greatly smoothing the process of learning the common feature space.

B. High-Order Structure Learning (HSL) Module

Suppose that we have paired VIS-IR images, denoted as $\{\mathbf{I}^{vis}, \mathbf{I}^{ir}\}$ with the same identity label. We first extract VIS features $\mathbf{B}^{vis}$ and IR features $\mathbf{B}^{ir}$ from the backbone network, respectively. Then, $\mathbf{B}^{vis}$ and $\mathbf{B}^{ir}$ are passed through the SLE module (more details of SLE can be seen in our previous work [12]) to extract both short-range features ($\mathbf{F}_S^{vis}$/$\mathbf{F}_S^{ir}$) and long-range features ($\mathbf{F}_L^{vis}$/$\mathbf{F}_L^{ir}$) for VIS and IR modalities. Thus, we can obtain a feature set $\mathcal{Q} = \{\mathbf{F}_L^{vis}, \mathbf{F}_S^{vis}, \mathbf{F}_L^{ir}, \mathbf{F}_S^{ir}\}$. The sizes of each feature in $\mathcal{Q}$ are $\mathbb{R}^{H\times W\times C}$, where H, W, and C correspond to the height, the width, and the number of channels of the features, respectively.

The backbone network and the SLE module just capture pixel-level and region-level dependencies within person images. However, they can not fully exploit the high-order structural information that delineates complex relationships among features (e.g., the head, torso, upper arm, and lower arm are parts of the upper body while head, torso, arm, and leg belong to the whole body). Inspired by the Hypergraph Neural Network (HGNN) [32], we introduce an HSL module to better capture high-order correlations, thereby enriching the feature representations. Besides, due to the small differences in the feature nodes of pedestrians, the conventional HGNN tends to suffer from the problem of model collapse that leads to the diverse and complex relationships tending to be the same. To deal with this problem, we make good use of the whitening operation and apply it to the hypergraph network,

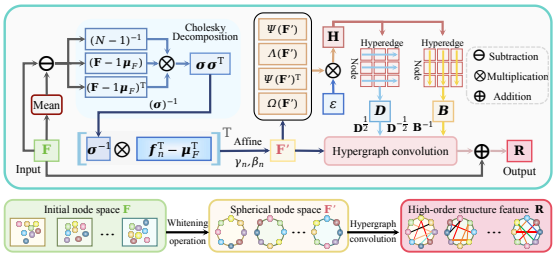


Fig. 3. The detailed network architecture of the proposed HSL module.

as shown in Fig. 3.

Different from the typical graph models that connect only pairwise nodes, hypergraphs provide a more sophisticated structure by allowing connections between an arbitrary number of nodes, thereby describing the high-order structural information. For each feature within the set $\mathcal{Q}$, we construct a whitened hypergraph, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$. Here, $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$, $\mathcal{E} = \{e_1, e_2, \cdots, e_M\}$ and $\mathbf{W}$ represent the node set, the hyperedge set and the weight matrix, respectively. $N = HW$ and M correspond to the numbers of nodes and hyperedges, respectively. In this paper, each $1 \times 1 \times C$ grid from the feature in $\mathcal{Q}$ is considered as a feature node. The n-th node is represented as $\mathbf{f}_n \in \mathbb{R}^{1\times C}$, and thus all nodes can be represented by $\mathbf{F} = [\mathbf{f}_1; \mathbf{f}_2; \cdots; \mathbf{f}_N] \in \mathbb{R}^{N\times C}$.

The conventional hypergraph network [32] is designed to enable unrestricted node connections to capture high-order structural information. However, it easily suffers from model collapse (i.e., the nodes connected by different hyperedges are the same) during hypergraph learning. To overcome this difficulty, we introduce a whitening operation to project the nodes into a spherical distribution and facilitate the learning of subtle high-order relationships. The whitening operation plays the role of "scattering" on the nodes, thereby preventing the diverse high-order connections from converging into a single connection. As a result, this approach enables us to explore various high-order relationships across these features effectively.

The whitened node $\mathbf{f}'_n$ can be obtained as

$$\mathbf{f}'_n = \gamma_n(\sigma^{-1}(\mathbf{f}_n^{\mathrm{T}} - \mu_{\mathbf{F}}^{\mathrm{T}}))^{\mathrm{T}} + \beta_n, \tag{1}$$

where $\sigma \in \mathbb{R}^{C\times C}$ denotes the lower triangular matrix that is obtained by the cholesky decomposition $\sigma\sigma^{\mathrm{T}} = \frac{1}{N-1}(\mathbf{F} - \mathbf{1}\mu_{\mathbf{F}})^{\mathrm{T}}(\mathbf{F} - \mathbf{1}\mu_{\mathbf{F}})$; $\mu_{\mathbf{F}} \in \mathbb{R}^{1\times C}$ denotes the mean vector of $\mathbf{F}$; $\mathbf{1} \in \mathbb{R}^{N\times 1}$ is a column vector of all ones; $\gamma_n \in \mathbb{R}^{1\times 1}$ and $\beta_n \in \mathbb{R}^{1\times C}$ are the learnable affine parameters. In such a way, $\mathbf{F}' = [\mathbf{f}'_1; \mathbf{f}'_2; \cdots; \mathbf{f}'_N] \in \mathbb{R}^{N\times C}$ can represent all the whitened nodes, where $\mathbf{f}_n \in \mathbb{R}^{1\times C}$ is the n-th node in $\mathbf{F}'$.

Similarly to [38], we use cross-correlation to learn the incidence matrix $\mathbf{H} \in \mathbb{R}^{N\times M}$, i.e.,

$$\mathbf{H} = \varepsilon(\Psi(\mathbf{F}')\Lambda(\mathbf{F}')\Psi(\mathbf{F}')^{\mathrm{T}}\Omega(\mathbf{F}')), \tag{2}$$

where $\Psi(\mathbf{F}') \in \mathbb{R}^{N\times C}$ introduces the learnable parameters to perform the linear transformation for the whitened nodes for all the whitened nodes $\mathbf{F}'$. $\Omega(\mathbf{F}') \in \mathbb{R}^{N\times M}$ gets M hyperedges of whitened nodes by trainable parameters for the $\mathbf{F}'$. The diagonal operation in $\Lambda(\cdot)$ is used to capture the context relationship of the whitened nodes and determine their distance contributions to the corresponding hyperedges, where $\Lambda(\mathbf{F}') \in \mathbb{R}^{C\times C}$. $\varepsilon(\cdot)$ is the step function. Hence in this end-to-end trainable way, the high-order structure information in person features can be well exploited with $\mathbf{H}$.

Based on the above $\mathbf{H}$, we introduce the hypergraph convolutional operation [32] to aggregate high-order structure information and the high-order relation-enhanced feature $\mathbf{R} \in \mathbb{R}^{N\times C}$ can be obtained as

$$\mathbf{R} = (\mathbf{I} - \mathbf{D}^{1/2}\mathbf{H}\mathbf{W}\mathbf{B}^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{D}^{-1/2})\mathbf{F}'\Theta + \mathbf{F}, \tag{3}$$

where $\mathbf{I} \in \mathbb{R}^{N\times N}$ is the identity matrix; $\mathbf{W} \in \mathbb{R}^{M\times M}$ denotes the weight matrix; $\mathbf{D} \in \mathbb{R}^{N\times N}$ and $\mathbf{B} \in \mathbb{R}^{M\times M}$ represent the node degree matrix and the hyperedge degree matrix obtained by the broadcast operation, respectively; $\Theta \in \mathbb{R}^{C\times C}$ denotes the learnable parameters. Following the above steps, we feed features from $\mathcal{Q}$ into the HSL module and obtain a relation-enhanced feature set $\mathcal{R} = \{\mathbf{R}_L^{vis}, \mathbf{R}_S^{vis}, \mathbf{R}_S^{ir}, \mathbf{R}_L^{ir}\}$, where each feature in $\mathcal{R}$ is obtained by Eq. (3).

C. Fine-Coarse Graph Attention Alignment (FCGA) Module

Some existing feature-level methods [1], [7] try to directly reduce the distances between VIS and IR features by the loss function, which can not achieve a reliable common feature space because of the large modality gap. Later, some methods [25], [26], [39] generate the auxiliary features from one or two modalities to mitigate the modality discrepancy, but they still lack efficient alignment and full utilization between different modality features. In order to effectively mitigate modality discrepancy, we leverage a fine-coarse graph attention alignment (FCGA) module, which aligns the features from different modalities and ranges by combining the fine-grain graph attention alignment (FGA) with the coarse-grain graph attention alignment (CGA), so as to generate reliable middle features, as shown in Fig. 4. In the FCGA module, the short-range features can offer local details to long-range features, making pedestrian feature representation more discriminative, while the long-range features can provide contextual information for short-range features to focus on the global relationship between detailed features.

During the feature alignment, the fine-grain graph attention establishes the dense connections between feature nodes that can reserve the details of middle features. Besides, the coarse-grain graph attention perceives the semantic associations of regional feature nodes to improve the quality of the overall middle feature. Specifically, we align each feature with the other three features in $\mathcal{R}$ and generate a middle feature, which involves the information from different modalities and ranges. We take the alignment between two features $\mathbf{R}_L^{vis}$ and $\mathbf{R}_S^{ir}$ as an example.

For the fine-grain graph attention alignment (FGA), first, we establish the similarity matrix $\mathbf{U} \in \mathbb{R}^{N\times N}$ between $\mathbf{R}_L^{vis}$ and $\mathbf{R}_S^{ir}$ by using the inner product and the softmax function, which can be formulated as

$$\mathbf{U} = \mathrm{Softmax}(\mathbf{R}_L^{vis}\theta_{q,f}(\mathbf{R}_S^{ir}\theta_{k,f})^{\mathrm{T}}), \tag{4}$$

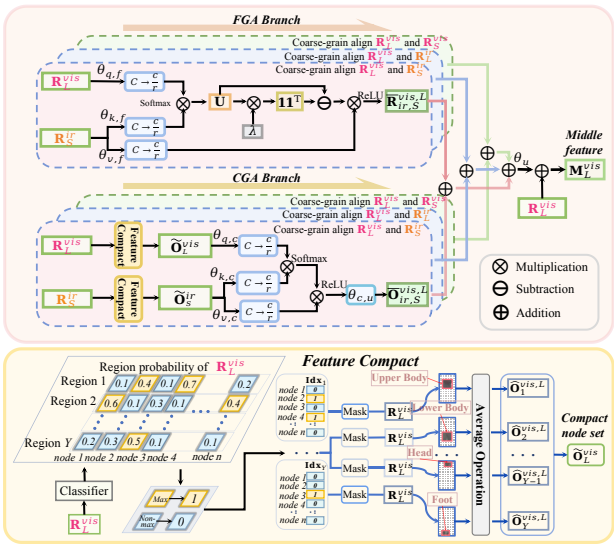


Fig. 4. The detailed network architecture of the proposed FCGA module (Taking aligning $\mathbf{R}_L^{vis}$ with $\mathbf{R}_S^{ir}$, $\mathbf{R}_L^{ir}$, and $\mathbf{R}_S^{vis}$ as an example).

where $\theta_{q,f} \in \mathbb{R}^{C\times\frac{C}{r}}$ and $\theta_{k,f} \in \mathbb{R}^{C\times\frac{C}{r}}$ are linear transformations; r is a reduction ratio used to perform the squeeze and excitation, thereby reducing the number of parameters; $\mathrm{Softmax}(\cdot)$ denotes the softmax function.

Then, we adopt the graph attention [37] to perform alignment between $\mathbf{R}_L^{vis}$ and $\mathbf{R}_S^{ir}$ according to the similarity matrix. Therefore, the aggregated node $\bar{\mathbf{R}}_{ir,S}^{vis,L} \in \mathbb{R}^{N\times\frac{C}{r}}$ is

$$\begin{aligned}\bar{\mathbf{R}}_{ir,S}^{vis,L} &= \mathrm{FGA}(\mathbf{R}_L^{vis}, \mathbf{R}_S^{ir}) \\ &= \mathrm{ReLU}(\mathbf{U} - \lambda\mathrm{Mean}(\mathbf{redU})\mathbf{1}\mathbf{1}^{\mathrm{T}})(\mathbf{R}_S^{ir}\theta_{v,f}),\end{aligned} \tag{5}$$

]where $\mathrm{FGA}(\cdot)$ denotes the fine-grain graph attention alignment operation; $\theta_{v,f} \in \mathbb{R}^{C\times\frac{C}{r}}$ is the linear transformation; λ is the balancing parameter that reduces nodes with low similarity; $\mathbf{1}\mathbf{1}^{\mathrm{T}} \in \mathbb{R}^{N\times N}$ is a matrix of all ones; and $\mathrm{ReLU}(\cdot)$ and $\mathrm{Mean}(\cdot)$ represent the ReLU activation function that sets similarity values less than 0 to 0 and the mean operation, respectively.

Different from the fine-grain alignment that needs to refine the details of each node-level feature, by compacting the feature nodes (e.g., classify each node and merge similar nodes into a single node), the representation ability of the feature nodes can be further enhanced, and the computational complexity of the feature alignment process can be reduced. So, we introduce a coarse-grain graph attention alignment (CGA) to improve the efficiency of generating intermediate features and enhance the semantic association of the middle features.

To begin with, we classify each feature node of $\mathbf{R}_L^{vis}$ into Y regions (i.e., head, arm, torso, leg, and so on) by a learnable classifier, and achieve the total region probability $\mathbf{P} \in \mathbb{R}^{N\times Y}$, that is,

$$\mathbf{P} = \mathrm{Classifier}(\mathbf{R}_L^{vis}) = [\mathbf{p}_1,\ \mathbf{p}_2,\ \cdots,\ \mathbf{p}_N], \tag{6}$$

where $\mathrm{Classifier}(\cdot)$ is a fully connected layer which consists of learnable parameters $\theta_{cls} \in \mathbb{R}^{C\times Y}$; $\mathbf{p}_n \in \mathbb{R}^{1\times Y}$ represents the probability values that the n-th $(n \in \{1, 2, \cdots, N\})$ feature node respectively belongs to the Y regions.

Then, we denote $idx_{n,z}$ as the index value of the n-th feature node belonging to the z-th $(z \in \{1, 2, \cdots, Y\})$ region. The $idx_{n,z}$ can be defined as

$$idx_{n,z} = \begin{cases} 1, & if \ \underset{y\in\{1,2,\cdots,Y\}}{\arg\max}\ (p_{n,y}) = z \\ 0, & otherwise \end{cases}, \tag{7}$$

where $\arg\max(\cdot)$ returns the index of the maximum probability value in $\mathbf{p}_n = [p_{n,1},\ p_{n,2},\ \cdots,\ p_{n,Y}] \in \mathbb{R}^{1\times Y}$, and $p_{n,y}$ means the probability value of the n-th feature node belonging to the y-th region.

For the z-th region, we can get the index vector $\mathbf{idx}_z = [idx_{1,z},\ idx_{2,z},\ \cdots,\ idx_{N,z}]$, and then the feature node set $\mathbf{O}_z^{vis,L}$ with the same z-th region can be obtained as

$$\mathbf{O}_z^{vis,L} = \mathrm{Mask}(\mathbf{R}_L^{vis}, \mathbf{idx}_z), \tag{8}$$

where $\mathrm{Mask}(\cdot)$ sets the value of the n-th feature node to 0 when the corresponding $idx_{n,z}$ is 0; otherwise keep the same value as before.

Finally, we perform an average operation on feature nodes $\mathbf{O}_z^{vis,L} = [\mathbf{o}_{z,1}^{vis,L}, \mathbf{o}_{z,2}^{vis,L}, \cdots, \mathbf{o}_{z,N}^{vis,L}] \in \mathbb{R}^{N\times C}$ to obtain the z-th compact representation $\hat{\mathbf{O}}_z^{vis,L}$, i.e.,

$$\hat{\mathbf{O}}_z^{vis,L} = \frac{\sum_{n=1}^{N} \mathbf{o}_{z,n}^{vis,L}}{N_z^{vis,L} + \epsilon}, \tag{9}$$

where $\mathbf{o}_{z,n}^{vis,L}$ represents the n-th feature node of $\mathbf{O}_z^{vis,L}$; $N_z^{vis,L}$ means the number of non-zero feature nodes in $\mathbf{O}_z^{vis,L}$; ϵ is a very small value (i.e., e^{-5}) added to the denominator for numerical stability. Similarly to Eq. (9), we can get the other Y-1 compact feature nodes and thus the final compact feature node set can be represented as $\tilde{\mathbf{O}}_L^{vis} \in \mathbb{R}^{Y\times C}$.

Moreover, to improve the efficiency of feature alignment, and explore contextual semantic associations in different modalities and ranges. Hence, we design a coarse-grain graph attention (CGA) to align $\tilde{\mathbf{O}}_L^{vis}$ and $\tilde{\mathbf{O}}_S^{ir}$, and the aligned compact feature node set $\bar{\mathbf{O}}_{ir,S}^{vis,L} \in \mathbb{R}^{N\times\frac{C}{r}}$ can be formulated as

$$\begin{aligned}\bar{\mathbf{O}}_{ir,S}^{vis,L} &= \mathrm{CGA}(\tilde{\mathbf{O}}_L^{vis}, \tilde{\mathbf{O}}_S^{ir}) \\ &= \theta_{c,u}\mathrm{ReLU}(\mathrm{Softmax}(\tilde{\mathbf{O}}_L^{vis}\theta_{q,c}(\tilde{\mathbf{O}}_S^{ir}\theta_{k,c})^{\mathrm{T}}))(\tilde{\mathbf{O}}_S^{ir}\theta_{v,c}),\end{aligned} \tag{10}$$

where $\theta_{q,c} \in \mathbb{R}^{C\times\frac{C}{r}}$, $\theta_{k,c} \in \mathbb{R}^{C\times\frac{C}{r}}$, $\theta_{v,c} \in \mathbb{R}^{C\times\frac{C}{r}}$ and $\theta_{c,u} \in \mathbb{R}^{N\times Y}$ are the linear transformations. Unlike FGA, the CGA allocates all feature nodes to the limited Y regions, without suppressing low similarity feature regions to ensure the efficiency of the coarse-grain alignment.

Based on the above, we employ an effective and efficient fine-coarse graph attention alignment (FCGA) module to align different modalities and ranges of feature nodes from node-level and region-level, respectively, as follows

$$\begin{aligned}\mathrm{FCGA}(\mathbf{R}_L^{vis}, \mathbf{R}_S^{ir}) = \mathrm{FGA}(\mathbf{R}_L^{vis}, \mathbf{R}_S^{ir}) + \\ \mathrm{CGA}(\mathbf{R}_L^{vis}, \mathbf{R}_S^{ir}).\end{aligned} \tag{11}$$

Through the FCGA module with a small number of parameters, a middle feature $\mathbf{M}_L^{vis} \in \mathbb{R}^{N\times C}$ can be obtained by aligning $\mathbf{R}_L^{vis}$ with the other three features $\mathbf{R}_S^{ir}$, $\mathbf{R}_L^{ir}$, and $\mathbf{R}_S^{vis}$ in $\mathbf{R}$, that is,

$$\begin{aligned}\mathbf{M}_L^{vis} = (&\mathrm{FCGA}(\mathbf{R}_L^{vis}, \mathbf{R}_S^{ir}) + \mathrm{FCGA}(\mathbf{R}_L^{vis}, \mathbf{R}_L^{ir}) + \\ &\mathrm{FCGA}(\mathbf{R}_L^{vis}, \mathbf{R}_S^{vis}))\theta_u + \mathbf{R}_L^{vis},\end{aligned} \tag{12}$$

where $\theta_u \in \mathbb{R}^{\frac{C}{r}\times C}$ represents the linear transformation. We can get the other reliable middle features similar to Eq. (12). Hence, we obtain the middle feature set $\tilde{\mathcal{R}} = \{\mathbf{M}_L^{vis}, \mathbf{M}_S^{vis}, \mathbf{M}_L^{ir}, \mathbf{M}_S^{ir}\}$.

D. Hierarchical Middle-Feature Agent Learning (HMAL) Loss

We generate agents of middle features at each modality-shared feature extraction stage to assist the network in learning a better common feature space and reducing the modality discrepancy hierarchically. For example, we generate the middle-feature agent $\mathbf{A}_1 \in \mathbb{R}^{1\times1\times C_1}$ for Stage 1 of the network based on the above middle feature set $\tilde{\mathcal{R}}$, that is,

$$\mathbf{A}_1 = \theta_{m,1}(\mathrm{Avg}(\tilde{\mathcal{R}})), \tag{13}$$

where $\mathrm{Avg}(\cdot)$ is the average pooling operation; $\theta_{m,1}$ is the learnable parameter that adjusts the feature size to $1\times1\times C_1$ (C_1 means the number of channels of the features in Stage 1).

Then, we introduce the modality discrepancy reduction (MDR) loss to reduce the difference between the pooled VIS feature ($\tilde{\mathbf{F}}_{1,vis}$), IR feature ($\tilde{\mathbf{F}}_{1,ir}$) and middle-feature agent ($\mathbf{A}_1$) in Stage 1 of the network, as follows

$$\mathcal{L}_{MDR}^{\tilde{\mathbf{F}}_1\leftrightarrow\mathbf{A}_1} = \mathrm{L}_1(\tilde{\mathbf{F}}_{1,vis} + \tilde{\mathbf{F}}_{1,ir}, 2\mathbf{A}_1) + \mathrm{L}_1(\tilde{\mathbf{F}}_{1,vis}, \tilde{\mathbf{F}}_{1,ir}). \tag{14}$$

where $\mathrm{L}_1(\cdot)$ represents the L1 distance; $\tilde{\mathbf{F}}_{1,vis} \in \mathbb{R}^{1\times1\times C_1}$ and $\tilde{\mathbf{F}}_{1,ir} \in \mathbb{R}^{1\times1\times C_1}$ mean the VIS feature $\mathbf{F}_{1,vis}$ and IR feature $\mathbf{F}_{1,ir}$ after the pooling operation, respectively.

Similarly to Eq. (13), we can achieve other middle-feature agents (i.e., $\mathbf{A}_2$, $\mathbf{A}_3$, and $\mathbf{A}_4$), and the total MDR loss can be expressed as:

$$\mathcal{L}_{HMAL}^{MDR} = \mathcal{L}_{MDR}^{\tilde{\mathbf{F}}_1\leftrightarrow\mathbf{A}_1} + \mathcal{L}_{MDR}^{\tilde{\mathbf{F}}_2\leftrightarrow\mathbf{A}_2} + \mathcal{L}_{MDR}^{\tilde{\mathbf{F}}_3\leftrightarrow\mathbf{A}_3} + \mathcal{L}_{MDR}^{\tilde{\mathbf{F}}_4\leftrightarrow\mathbf{A}_4}. \tag{15}$$

With the agent of the middle features, our method can learn discriminative features from all the network stages and hierarchically enhance feature representations to achieve a reliable common feature space between different modalities.

The existing VI-ReID methods [1], [40] follow low-to-high feature extraction, which ignores the interaction of features at different stages. The features of the low stage contain more detailed information, while the features of the high stage have rich semantic relationships. In this subsection, we will use the BFE loss to build the mutual interaction bridge and thus achieve the bi-directional enhancement between the features of different stages. The interactions of the high-to-low and the low-to-high can improve the ability of the network to capture discriminative features. In other words, the low-stage network can focus on detailed features guided by semantic relationships from the high-stage, and the high-stage network can enhance the semantic relationship by using detailed information from the low-stage.

We take the bi-directional enhancement between the features of Stage 1 (S_1) and Stage 2 (S_2) as an example. First, we adopt the pooling operation on the features of S_1 and S_2 to obtain the feature representations $\tilde{\mathbf{F}}_1 \in \mathbb{R}^{1\times1\times C_1}$ and $\tilde{\mathbf{F}}_2 \in \mathbb{R}^{1\times1\times C_2}$ (C_2 is the number of channels of the features in S_2), that is, $\tilde{\mathbf{F}}_1 = \mathrm{Avg}(\mathbf{F}_1)$ and $\tilde{\mathbf{F}}_2 = \mathrm{Avg}(\mathbf{F}_2)$. Here $\mathrm{Avg}(\cdot)$ is the average pooling operation; $\mathbf{F}_1 \in \mathbb{R}^{H_1\times W_1\times C_1}$ and $\mathbf{F}_2 \in \mathbb{R}^{H_2\times W_2\times C_2}$ are the features of S_1 and S_2, respectively (H_1/H_2 and W_1/W_2 correspond to the height and width of the features in S_1/S_2, respectively).

Then, we perform the upsample and downsample operations on the features $\tilde{\mathbf{F}}_1$ and $\tilde{\mathbf{F}}_2$ to achieve $\tilde{\mathbf{F}}_1^{up} \in \mathbb{R}^{1\times1\times C_2}$ and $\tilde{\mathbf{F}}_2^{down} \in \mathbb{R}^{1\times1\times C_1}$, respectively, which can be formulated as

$$\tilde{\mathbf{F}}_1^{up} = \mathrm{Upsample}(\tilde{\mathbf{F}}_1); \quad \tilde{\mathbf{F}}_2^{down} = \mathrm{Downsample}(\tilde{\mathbf{F}}_2), \tag{16}$$

where the $\mathrm{Upsample}(\cdot)$ and $\mathrm{Downsample}(\cdot)$ operations make $\tilde{\mathbf{F}}_1^{up}$ and $\tilde{\mathbf{F}}_2^{down}$ become the same size as $\mathbf{F}_2$ and $\mathbf{F}_1$ by linear transformations.

To provide detailed information from low-stage features to high-stage features and transfer high-stage semantic information to low-stage features, we adopt the L_1 distance to perform the bi-directional interaction between the features of S_1 and S_2, which is defined as

$$\mathcal{L}_{BFE}^{S_1\leftrightarrow S_2} = \mathrm{L}_1(\tilde{\mathbf{F}}_1^{up}, \tilde{\mathbf{F}}_2) + \mathrm{L}_1(\tilde{\mathbf{F}}_2^{down}, \tilde{\mathbf{F}}_1). \tag{17}$$

We also perform bi-directional enhancement between S_2 and S_3 and between S_3 and S_4, and thus the final BFE loss can be written as

$$\mathcal{L}_{HMAL}^{BFE} = \mathcal{L}_{BFE}^{S_1\leftrightarrow S_2} + \mathcal{L}_{BFE}^{S_2\leftrightarrow S_3} + \mathcal{L}_{BFE}^{S_3\leftrightarrow S_4}. \tag{18}$$

The final HMAL loss, which is defined as

$$\mathcal{L}_{HMAL} = \mathcal{L}_{HMAL}^{MDR} + \mathcal{L}_{HMAL}^{BFE}. \tag{19}$$

E. Modality-Range Identity-Center Contrastive (MRIC) Loss

To reduce the intra-class difference and increase inter-class discrepancy, we introduce the MRIC loss to improve feature representations and minimize the modality gaps among the VIS, IR, and middle features. The MRIC loss consists of three items: an intra-range loss, a middle feature loss, and an inter-modality loss based on identity centers. The illustration of the MRIC loss is presented in Fig. 5.

Following previous works [1], [7], we apply the holistic and partial generalized mean pooling to each feature in $\tilde{\mathcal{R}}$ and concatenate the pooling features to obtain the 1D middle features, and we can get the 1D middle feature set $\tilde{\mathcal{R}}' = \{\mathbf{m}_L^{vis}, \mathbf{m}_S^{vis}, \mathbf{m}_L^{ir}, \mathbf{m}_S^{ir}\}$. Analogously, we apply the same pooling and concatenation operations to each feature in $\mathcal{R}$ and thus obtain the 1D feature set $\mathcal{R}' = \{\mathbf{r}_L^{vis}, \mathbf{r}_S^{vis}, \mathbf{r}_L^{ir}, \mathbf{r}_S^{ir}\}$.

The robustness of the identity centers ensures they are not influenced by pedestrian appearance changes. Technically, we first obtain identity centers through the weighted average of the features of each person at the specific modality and range. For example, the center of the relation-enhanced features for the

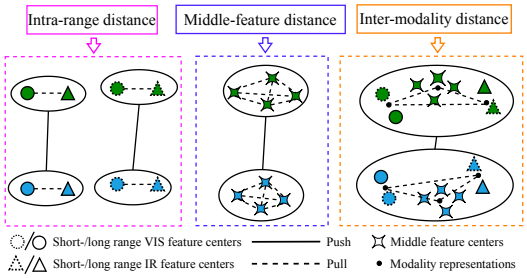


Fig. 5. Illustration of the proposed MRIC loss. Different colors represent different identities.

pedestrian with the identity i at the long-range VIS modality can be achieved as

$$\mathbf{c}_{L,i}^{vis} = \sum_{j=1}^{K} \frac{\exp(\sum_{k=1}^{K} \mathbf{r}_{L,i,j}^{vis} \mathbf{r}_{L,i,k}^{vis\ \mathrm{T}})}{\sum_{j=1}^{K} \exp(\sum_{k=1}^{K} \mathbf{r}_{L,i,j}^{vis} \mathbf{r}_{L,i,k}^{vis\ \mathrm{T}})} \mathbf{r}_{L,i,j}^{vis}, \tag{20}$$

where K represents the number of VIS features of each person; $\mathbf{r}_{L,i,k}^{vis} \in \mathbb{R}^{1\times C'}$ denotes the k-th 1D relation-enhanced long-range VIS feature with the identity i in $\mathcal{R}'$.

Accordingly, we can obtain the identity center sets $\mathcal{C}_L^{vis}$ $(\{\mathbf{c}_{L,i}^{vis}\}_{i=1}^{P})$, $\mathcal{C}_S^{vis}$ $(\{\mathbf{c}_{S,i}^{vis}\}_{i=1}^{P})$, $\mathcal{C}_L^{ir}$ $(\{\mathbf{c}_{L,i}^{ir}\}_{i=1}^{P})$, $\mathcal{C}_S^{ir}$ $(\{\mathbf{c}_{S,i}^{ir}\}_{i=1}^{P})$, $\tilde{\mathcal{C}}_L^{vis}$ $(\{\tilde{\mathbf{c}}_{L,i}^{vis}\}_{i=1}^{P})$, $\tilde{\mathcal{C}}_S^{vis}$ $(\{\tilde{\mathbf{c}}_{S,i}^{vis}\}_{i=1}^{P})$, $\tilde{\mathcal{C}}_L^{ir}$ $(\{\tilde{\mathbf{c}}_{L,i}^{ir}\}_{i=1}^{P})$, and $\tilde{\mathcal{C}}_S^{ir}$ $(\{\tilde{\mathbf{c}}_{S,i}^{ir}\}_{i=1}^{P})$, where $\mathcal{C}$ and $\tilde{\mathcal{C}}$ represent the center sets for the enhanced features and the middle features at a specific range and modality, respectively; P is the number of pedestrian identities in the training set.

The intra-range loss $\mathcal{L}_{MRIC}^{SL}$ is to reduce the distances between the same-range VIS and IR features from the same pedestrian while enlarging the distances between the same-range VIS and IR features from different pedestrian, that is,

$$\mathcal{L}_{MRIC}^{SL} = \mathcal{L}_{MRIC}^{\mathcal{C}_S^{vis},\mathcal{C}_S^{ir}} + \mathcal{L}_{MRIC}^{\mathcal{C}_L^{vis},\mathcal{C}_L^{ir}}, \tag{21}$$

where

$$\begin{aligned}\mathcal{L}_{MRIC}^{\mathcal{A},\mathcal{B}} = &-\sum_{i=1}^{P} \log \frac{\exp(\mathcal{M}_{i,i}^{\mathcal{A},\mathcal{B}})}{\sum_{z=1}^{P} \exp(\mathcal{M}_{i,z}^{\mathcal{A},\mathcal{B}})} \\ &-\sum_{i=1}^{P} \log \frac{\exp(\mathcal{M}_{i,i}^{\mathcal{A},\mathcal{B}})}{\sum_{z=1}^{P} \exp(\mathcal{M}_{z,i}^{\mathcal{A},\mathcal{B}})} \\ &+\sum_{i=1}^{P} \mathrm{L}_1(\mathbf{a}_i - \mathbf{b}_i).\end{aligned} \tag{22}$$

Here, $\mathcal{M}^{\mathcal{A},\mathcal{B}} \in \mathbb{R}^{P\times P}$ denotes the cosine similarity matrix between $\mathcal{A}$ and $\mathcal{B}$. $\mathcal{M}_{i,j}^{\mathcal{A},\mathcal{B}}$ denotes the cosine similarity between the i-th row $(\mathbf{a}_i)$ of matrix $\mathcal{A}$ and the j-th row $(\mathbf{b}_j)$ of matrix $\mathcal{B}$; $\mathrm{L}_1(\cdot)$ represents the L_1 norm. By minimizing the $\mathcal{L}_{MRIC}^{\mathcal{A},\mathcal{B}}$, we can effectively decrease and increase the distance between the same pedestrian and different pedestrians in the feature space, respectively.

The middle-feature loss $\mathcal{L}_{MRIC}^{MID}$ reduces the distances between different middle features, which is defined as

$$\begin{aligned}\mathcal{L}_{MRIC}^{MID} = &\mathcal{L}_{MRIC}^{\tilde{\mathcal{C}}_S^{vis},\tilde{\mathcal{C}}_L^{vis}} + \mathcal{L}_{MRIC}^{\tilde{\mathcal{C}}_S^{vis},\tilde{\mathcal{C}}_S^{ir}} + \mathcal{L}_{MRIC}^{\tilde{\mathcal{C}}_S^{vis},\tilde{\mathcal{C}}_L^{ir}} + \\ &\mathcal{L}_{MRIC}^{\tilde{\mathcal{C}}_L^{vis},\tilde{\mathcal{C}}_S^{ir}} + \mathcal{L}_{MRIC}^{\tilde{\mathcal{C}}_L^{vis},\tilde{\mathcal{C}}_L^{ir}} + \mathcal{L}_{MRIC}^{\tilde{\mathcal{C}}_S^{ir},\tilde{\mathcal{C}}_L^{ir}}.\end{aligned} \tag{23}$$

The inter-modality loss $\mathcal{L}_{MRIC}^{VI}$ is leveraged to mitigate the intra-class discrepancy and enlarge the inter-class distances between VIS, IR, and middle features, which is formulated as

$$\mathcal{L}_{MRIC}^{VIM} = \mathcal{L}_{MRIC}^{\mathcal{C}^{vis},\mathcal{C}^{ir}} + \mathcal{L}_{MRIC}^{\mathcal{C}^{vis},\mathcal{C}^{mid}} + \mathcal{L}_{MRIC}^{\mathcal{C}^{ir},\mathcal{C}^{mid}}, \tag{24}$$

where $\mathcal{C}^{vis}$, $\mathcal{C}^{ir}$, and $\mathcal{C}^{mid}$ represent the identity center sets of VIS, IR, and middle features, respectively; $\mathcal{C}^{vis}$ and $\mathcal{C}^{ir}$ denote the averaged features from the same modality for each person; $\mathcal{C}^{mid}$ is obtained by averaging all the middle features for each person. Thus, the MRIC loss is

$$\mathcal{L}_{MRIC} = \mathcal{L}_{MRIC}^{SL} + \mathcal{L}_{MRIC}^{MID} + \mathcal{L}_{MRIC}^{VIM}. \tag{25}$$

Finally, we adopt the cross-entropy loss ($\mathcal{L}_{CE}$ [41]), the triplet loss ($\mathcal{L}_{TRI}$ [42]), the HMAL loss ($\mathcal{L}_{HMAL}$), and the MRIC loss ($\mathcal{L}_{MRIC}$) to jointly train the HOH-Net. The joint loss $\mathcal{L}$ is defined as

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{TRI} + \mathcal{L}_{HMAL} + \mathcal{L}_{MRIC}. \tag{26}$$

IV. Experiments

A. Experimental Settings

Image-based Datasets. The SYSU-MM01 [22] dataset contains 491 identities. Its training set includes 395 identities with 22,258 VIS and 11,909 IR images, while the test set has 96 identities with 301 VIS and 3,803 IR images. The RegDB [43] dataset consists of 412 identities, each with 10 VIS and 10 IR images captured by two overlapping cameras. The LLCM [25] dataset provides 713 identities in the training set and 351 identities in the test set.

Video-based Dataset. The HITSZ-VCM dataset [44] is captured by 12 RGB and 12 IR cameras. Its training set includes 500 identities with 11,061 tracklets, while the test set contains 427 identities with 10,802 tracklets.

Implementation Details. During the training phase, all images are resized to $3 \times 288 \times 144$ with data augmentation [45]. For each mini-batch, we randomly select 8 identities with 4 VIS images and 4 IR images for each identity. We adopt AGW [1] as our backbone network. The learning rate is warmed up from 0.01 to 0.1 over the first 10 epochs, then decays to 0.01 at epoch 20 and 0.001 at epoch 50. We use SGD as the optimizer with a momentum parameter set to 0.9. The number of hyperedges M in the HSL module is set to 256. In the FCGA module, the reduction ratio r is set to 32. λ in Eq. (5) is set to 1.3, 1.1, 1.3, and 1.3 on the SYSU-MM01, RegDB, LLCM, and HITSZ-VCM datasets, respectively. In the FCGA module, the number Y of person regions is set to 9, 8, 9, and 9 on the SYSU-MM01, RegDB, LLCM, and HITSZ-VCM datasets, respectively. For HITSZ-VCM, each video sequence consists of 14 frames, averaged as the video representation. We train the HOH-Net for 120 epochs. The proposed HOH-Net is implemented in PyTorch on an NVIDIA A40 GPU.

B. Comparison with State-of-the-Art Methods

Our proposed HOH-Net is compared with some SOTA image-based and video-based models, including LbA [39], TSME [5], SPOT [20], DFLN-ViT [10], PMT [8], CAL [46],

TABLE I
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE SYSU-MM01, REGDB, AND LLCM DATASETS. * AND † REPRESENT THE IMAGE-LEVEL AND FEATURE-LEVEL METHODS, RESPECTIVELY. R-1 (%), MAP (%), FLOPS (G), AND PARAMS (M) ARE REPORTED. THE BOLD FONT AND THE UNDERLINE DENOTE THE BEST AND SECOND-BEST PERFORMANCE, RESPECTIVELY.

Methods	Venue	SYSU-MM01		RegDB		LLCM		FLOPs	Params
		All search	Indoor search	VIS to IR	IR to VIS	VIS to IR	IR to VIS		
		R-1 / mAP	R-1 / mAP	R-1 / mAP	R-1 / mAP	R-1 / mAP	R-1 / mAP		
LbA† [39]	ICCV'21	55.4 / 54.1	58.5 / 66.3	74.2 / 67.6	67.5 / 72.4	50.8 / 55.6	43.8 / 53.1	5.2	24.3
TSME* [5]	TCSVT'22	64.2 / 61.2	64.8 / 71.5	87.4 / 76.9	86.4 / 75.7	- / -	- / -	-	-
SPOT† [20]	TIP'22	65.3 / 62.3	69.4 / 74.6	80.4 / 72.5	79.4 / 72.3	- / -	- / -	-	-
DFLN-ViT† [10]	TMM'23	59.8 / 57.7	62.1 / 76.0	92.1 / 82.1	91.2 / 81.6	- / -	- / -	-	-
PMT† [8]	AAAI'23	67.5 / 65.0	71.7 / 76.5	84.8 / 76.6	84.2 / 75.1	58.4 / 62.1	49.9 / 57.2	18.1	85.6
CAL† [46]	ICCV'23	74.7 / 71.7	79.7 / 83.7	94.5 / 88.7	93.6 / 87.6	- / -	- / -	-	-
DEEN† [25]	CVPR'23	74.7 / 71.8	80.3 / 83.3	91.1 / 85.1	89.5 / 83.4	62.5 / 65.8	54.9 / 62.9	16.2	41.2
CSMSF† [28]	TMM'24	70.6 / 67.5	76.0 / 80.2	85.3 / 76.4	83.9 / 75.2	- / -	- / -	-	-
CAJ$_+$* [47]	TPAMI'24	71.5 / 68.2	76.0 / 78.4	85.7 / 79.7	84.9 / 78.6	- / -	- / -	5.2	**23.5**
EIFJLF * [15]	TCSVT'24	72.2 / 72.8	81.8 / 84.2	82.0 / 82.5	82.4 / 83.2	- / -	- / -	-	-
CSC-Net † [27]	TCSVT'24	72.7 / 69.6	78.6 / 82.1	91.0 / 86.4	89.4 / 85.7	- / -	- / -	-	-
DCPLNet † [48]	TII'24	74.0 / 70.4	78.3 / 81.9	94.3 / 87.3	91.7 / 84.8	60.5 / 63.2	53.4 / 59.8	-	-
DMPF † [49]	TNNLS'24	**76.4** / 71.6	82.3 / 84.9	88.8 / 81.0	88.9 / 81.9	- / -	- / -	-	-
AGPI^2* [50]	TIFS'25	72.2 / 70.6	83.5 / 84.3	89.0 / 83.9	87.9 / 83.0	- / -	- / -	-	-
CSCL† [24]	TMM'25	75.7 / 72.1	80.8 / 83.6	92.2 / 84.3	89.7 / 85.1	- / -	- / -	-	-
MDANet† [23]	TMM'25	75.8 / 73.0	80.1 / 81.8	92.4 / 82.7	91.8 / 81.9	- / -	- / -	**3.6**	25.5
HOS-Net† [12]	AAAI'24	75.6 / 74.2	84.2 / 86.7	94.7 / 90.4	93.3 / 89.2	64.9 / 67.9	56.4 / 63.2	14.3	83.4
HOH-Net† (Ours)	-	76.2 / **74.5**	**84.4 / 87.2**	**95.1 / 90.7**	**93.7 / 89.5**	**65.7 / 68.3**	**56.8 / 63.5**	11.5	58.8

TABLE II
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE HITSZ-VCM DATASET. * AND † REPRESENT THE IMAGE-LEVEL AND FEATURE-LEVEL METHODS, RESPECTIVELY. R-1 (%) AND MAP (%) ARE REPORTED. THE BOLD FONT AND THE UNDERLINE DENOTE THE BEST AND SECOND-BEST PERFORMANCE, RESPECTIVELY.

Methods	Venue	HITSZ-VCM	
		VIS to IR	IR to VIS
		R-1 / mAP	R-1 / mAP
AuxNet* [51]	TIFS'23	54.6 / 48.7	51.1 / 46.0
MITML† [44]	CVPR'22	64.5 / 47.7	63.7 / 45.3
IBAN* [14]	TCSVT'23	69.6 / 51.0	65.0 / 48.8
DMA† [52]	SPL'24	69.9 / 52.3	66.6 / 50.2
SAADG* [53]	ACM MM'23	73.1 / 56.1	69.2 / 53.8
CST† [54]	TMM'24	72.6 / 53.0	69.4 / 51.2
HOS-Net† [12]	AAAI'24	73.6 / 56.5	70.4 / 54.2
HOH-Net† (Ours)	-	**74.8 / 57.1**	**71.4 / 54.9**

TABLE III
THE INFLUENCE OF KEY COMPONENTS OF THE PROPOSED HOH-NET ON THE SYSU-MM01 AND REGDB DATASETS. R-1 (%), MAP (%), FLOPS (G), PARAMS (M), AND INFERENCE TIME (S) ON THE SYSU-MM01 DATASET ARE REPORTED.

#	Settings						SYSU-MM01	RegDB	Params / FOPLs / Inference
	SLE	HSL	FCGA	$\mathcal{L}_{HMAL}^{MDR}$	$\mathcal{L}_{HMAL}^{BFE}$	$\mathcal{L}_{MRIC}$	R-1 / mAP	R-1 / mAP	
1	-	-	-	-	-	-	69.9 / 66.9	85.0 / 79.1	5.2 / 23.5 / 69.3
2	✓	-	-	-	-	-	71.7 / 69.4	89.6 / 84.8	7.1 / 38.8 / 80.5
3	✓	✓	-	-	-	-	73.3 / 72.4	92.0 / 87.1	9.2 / 51.9 / 82.7
4	✓	-	✓	-	-	-	72.3 / 70.6	91.8 / 86.8	7.8 / 39.8 / 71.5
5	✓	✓	✓	-	-	-	74.1 / 72.7	92.5 / 88.0	9.9 / 52.9 / 82.7
6	✓	✓	✓	✓	-	-	74.5 / 73.0	93.8 / 89.1	10.0 / 53.3 / 82.7
7	✓	✓	✓	-	✓	-	74.8 / 73.3	94.2 / 89.7	11.4 / 58.4 / 82.7
8	✓	✓	✓	✓	✓	-	75.5 / 73.9	94.5 / 90.2	11.5 / 58.8 / 82.7
9	✓	✓	✓	-	-	✓	75.2 / 74.0	94.8 / 90.4	7.8 / 39.8 / 82.7
10	✓	✓	✓	✓	✓	✓	**76.2 / 74.5**	**95.1 / 90.7**	11.5 / 58.8 / 82.7

DEEN [25], DARD [55], CSMSF [28], CAJ$_+$ [47], EIFJLF [15], CSC-Net [27], DMPF [49], AGPI^2 [50], CSCL [24], MDANet [23], HOS-Net [12], AuxNet [51], MITML [44], IBAN [14], DMA [52], SAADG [53], and CST [54], and the comparison results on the three challenging image-based VI-ReID datasets (i.e., SYSU-MM01, RegDB, and LLCM) and one video-based VI-ReID dataset (i.e., HITSZ-VCM) are given in Tables I and II.

SYSU-MM01. Our proposed HOH-Net obtains the impressive performance among all the state-of-the-art methods on the SYSU-MM01 dataset, as shown in Table I. Specifically, compared to the image-level TSME [5], EIFJLF [15], and AGPI^2 [50] methods, the HOH-Net has achieved at least 3.9% increase in mAP for the all search mode. Compared with the previous feature-level CNN-based method LbA [39] and the Transformer-based method PMT [8], the HOH-Net surpasses them by at least 8.7%/9.5% in Rank-1/mAP for the all search mode. The HOH-Net introduces the bi-directional feature enhancement between different stages and the agent of middle features, so it outperforms our previous method HOS-Net [12] by 0.6% in Rank-1 for the all search mode, which shows that our proposed HOH-Net can more effectively improve the representation of features and better reduce the modality discrepancy.

RegDB. As shown in Table I, it is evident that the proposed HOH-Net achieves the best performance for the VIS to IR and the IR to VIS search modes. For the VIS to IR search mode, compared with the feature-level SPOT [20] method that relies on high-quality human body structure labels to extract modality-shared features, the HOH-Net improves Rank-1 and mAP by at least 14.7% and 18.2%, respectively. For the IR to VIS search mode, because the importance of high-order structure modeling is exploited, our HOH-Net outperformsthe feature-level MDANet [23] method by 7.6% in mAP, respectively. Besides, compared to the feature-level CNN-Transformer hybrid methods DFLN-ViT [10] and SPOT [20], our method improves the mAP by at least 7.9% for the IR to VIS search mode. This further demonstrates the superiority of our network, which is based on high-order structures and person structure label-free for VI-ReID.

LLCM. We also report the comparison results on the third challenging dataset LLCM in Table I. For the IR to VIS search mode, our HOH-Net outperforms the feature-level method DCPLNet [48] by 3.4%/3.7% in Rank-1/mAP. Moreover, the

HOH-Net performs significantly better than the embedding feature expansion network (e.g., DEEN [25]) for the VIS to IR search mode, achieving the best results with 65.7%/68.3% in Rank-1/mAP, respectively, which shows that the HOH-Net can learn a discriminative and reliable common feature space to narrow the gap.

HITSZ-VCM. Our proposed method can also be well applied to video-based VI-ReID field, and the results are shown in Table II. We adopt the MITML [44] as the video-based VI-ReID baseline. Compared to the image-level method IBAN [14] which learns modality-irrelevant features by utilizing anaglyph data from middle pedestrian images, the HOH-Net improves Rank-1 by 5.2% for the VIS to IR search mode. Compared to the feature-level method CST [54], our method improves the 3.7% mAP for the IR to VIS search mode, indicating that our method can effectively model high-order structure information of pedestrians and better reduce the modality gap. In addition, compared to the previous HOS-Net, the HOH-Net outperforms the previous it by 1.2% Rank-1 for the VIS to IR search mode. These results indicate that the HOH-Net can adopt the HMAL loss to generate middle-feature agents and perform the bi-directional feature enhancement to hierarchically mitigate the modality gap between VIS and IR video frames, thus improving the video-based VI-ReID performance.

Model Complexity. The efficiency comparison with other state-of-the-art methods is also given in Table I. Compared with PMT [8], our HOH-Net method reduces FLOPs/Params by 6.6G/26.8M, and the mAP of the indoor search mode increased by 9.5% on the SYSU-MM01 dataset. These experiments demonstrate that the proposed HOH-Net can reduce the modality differences between VIS and IR images more effectively by the light-weight fine-coarse graph attention alignment module and the hierarchical middle-feature agent learning. In addition, compared with our previous HOS-Net [12] method, the number of parameters of the HOH-Net is reduced by 29.5%, and the Rank-1 of the VIS to IR search mode is improved by 0.8% on the LLCM dataset.

C. Ablation Studies

Effectiveness of Key Components. As shown in Table III, we conduct ablation studies to validate the effectiveness of each key component of the proposed HOH-Net (including HSL, FCGA, $\mathcal{L}_{HMAL}$, and $\mathcal{L}_{MRIC}$). #1 represents the Baseline [1] method.

SLE: By incorporating the SLE (+1.9M, +15.3G) into the Baseline, #2 outperforms #1, achieving about 2.5% increase in mAP on the SYSU-MM01 dataset. This demonstrates the effectiveness of our SLE, leveraging both CNN and Transformer, to extract both the short- and long-range person features.

HSL: By introducing HSL (+2.1M, +13.1G), #3 further improves 3.0% increase in mAP on the SYSU-MM01 dataset, compared with #2. This verifies the effectiveness of HSL adopting a whitened hypergraph network to effectively model the high-order relationships between different feature nodes.

FCGA: By introducing FCGA (+0.7M, +1.0G) to #3, #5 has improved accuracy by 0.8% in Rank-1 on the SYSU-MM01 dataset. This demonstrates that it is effective to reduce the modality discrepancy by learning the reliable middle features via the fine-coarse graph attention alignment.

$\mathcal{L}_{HMAL}$**:** Inserting modality discrepancy reduction loss $\mathcal{L}_{HMAL}^{MDR}$ (+0.1M, +0.4G) and bi-directional feature enhancement loss $\mathcal{L}_{HMAL}^{BFE}$ (+1.5M, +5.5G) at #5 yields Rank-1 gains of 0.4% and 0.7% on the SYSU-MM01 dataset, respectively. #8 simultaneously adopts $\mathcal{L}_{HMAL}^{MDR}$ and $\mathcal{L}_{HMAL}^{BFE}$ to hierarchically mitigate the modality gap and improve the feature representation. Compared with #5, the Rank-1 performance of #8 has been improved by 2.0% on the RegDB dataset.

$\mathcal{L}_{MRIC}$**:** When we apply the MRIC loss (#10) to train the network, we achieve the best performance results (i.e., the a remarkable 74.5% and 90.7% mAP on the SYSU-MM01 and RegDB datasets, respectively). These results demonstrate the effectiveness of the MRIC loss that can reduce discrepancies between the VIS and IR modalities, obtaining a discriminative and reliable common embedding space.

Though, compared to the Baseline, our method increases the computational complexity (+6.3M) and has a higher number of parameters (+35.3G), its performance gains are significantly greater (e.g., our HOH-Net outperforms the Baseline by 7.6% in mAP on the SYSU-MM01 dataset). Additionally, the inference time of the HOH-Net is about 1.2 times longer than that of the Baseline, yet it remains acceptable for real-world applications. Though gradually adding modules (i.e., SLE, HSL, FCGA, and loss functions ($\mathcal{L}_{HMAL}$ and $\mathcal{L}_{MRIC}$)) slightly increase complexity, they also significantly improve the mAP from 79.1% to 90.7% on the RegDB dataset. It is essential to highlight that the FCGA module generates reliable middle features and middle-feature agents through the ground-truth labels of VIS and IR images at the stage of training. During the inference stage, the FCGA module and the loss functions are not used for cross-modality retrieval, because the label information is unavailable. Consequently, the inference time of variants with and without the FCGA module and the loss functions is the same.

TABLE IV
THE INFLUENCE OF SLE MODULE AND THE DIFFERENT NODE RELATIONSHIP MODELING METHODS ON THE SYSU-MM01 AND REGDB DATASETS. R-1 (%) AND MAP (%) ARE REPORTED.

Settings	SYSU-MM01	RegDB
	R-1 / mAP	R-1 / mAP
Baseline	69.9 / 66.9	85.0 / 79.1
Baseline+SLE w/o TB	70.6 / 66.9	87.2 / 81.4
Baseline+SLE w/o CB	71.0 / 68.4	88.1 / 83.0
Baseline+SLE	71.7 / 69.4	89.6 / 84.8
+PConv [56]	71.9 / 69.2	90.2 / 85.1
+DEConv [57]	72.2 / 70.6	90.7 / 85.5
+Vision-RWKV [58]	72.3 / 70.4	90.5 / 85.0
+GNN [59]	72.5 / 71.3	91.5 / 86.4
+Hypergraph [32]	72.5 / 70.3	91.1 / 86.3
+MambaVision [60]	72.7 / 71.9	91.4 / 86.7
+HSL (PCA)	72.6 / 71.0	91.2 / 86.6
+HSL (ZCA)	73.1 / 71.9	91.7 / 86.8
+HSL (Cholesky)	**73.3 / 72.4**	**92.0 / 87.1**

Influence of Convolutional Blocks and Transformer Blocks in the SLE Module. We compare models trained by using the

SLE without Transformer Blocks, denoted as "Baseline+SLE w/o TB", the SLE without Convolutional Blocks, denoted as "Baseline+SLE w/o CB", and the whole SLE, to evaluate the influence of the SLE module, as shown in Table IV. We adopt 3 convolutional blocks and 2 Transformer blocks to achieve different ranges of features as our previous work [12]. From Table IV, we can observe that when convolutional blocks and Transformer blocks are used in the SLE module, it achieves 71.7%/69.4% and 89.6%/84.8% in Rank-1 and mAP on these two datasets, respectively. Especially, the SLE module surpasses the "Baseline+SLE w/o TB" and "Baseline+SLE w/o CB" by 1.1% and 0.7% in Rank-1 on the SYSU-MM01 dataset, respectively. This indicates that the SLE can effectively explore different ranges of person features by combining CNN with Transformer.

Effectiveness of the Different Node Relationship Modeling Methods. To verify the effectiveness of our proposed HSL module with Cholesky, we compare it with the different node relationship modeling methods, i.e., PConv [56], DEConv [57], Vision-RWKV [58], GNN [59], Hypergraph [32], and MambaVision [60] under the setting of "Baseline+SLE [12]". The PConv and DEConv methods adopt the convolution kernel layer to extract features, and the GNN method models node relationships by discovering fixed nearest neighbor feature nodes. Different from the PConv, DEConv, and GNN that only focus on the limited node relationship modeling, our proposed HSL (Cholesky) module provides a more sophisticated structure by allowing connections between an arbitrary number of whitened nodes. As shown in Table IV, our proposed HSL (Cholesky) outperforms the DEConv by 1.8% in mAP on the SYSU-MM01 dataset. Compared to the long-range spatial-channel mix-based Vision-RWKV [58] and the state space model-based MambaVision [60], the HSL (Cholesky) method outperforms them by 1.0% and 0.6% in Rank-1 on the SYSU-MM01 dataset, respectively. Especially, compared to the GNN, the proposed HSL (Cholesky) brings 1.1% and 0.7% improvements in mAP on the SYSU-MM01 and RegDB datasets, respectively. It is important to note that the original hypergraph [32] allows unrestricted connections between nodes to capture high-order structural information, but it might suffer from the influence of model collapse during the hypergraph learning. We also analyze the impact of different whitening methods (i.e., PCA, ZCA, and Cholesky) on hypergraphs. Compared to hypergraph [32], our proposed HSL with PCA/ZCA/Cholesky can explore high-order relationships and achieves 71.0% (+0.7%), 71.9% (+1.6%), and 72.4% (2.1%) in mAP on the SYSU-MM01 dataset, respectively. This indicates that our HSL module can effectively model the complex and diverse high-order structure relationships between pedestrians, and can avoid the model collapse (i.e., the nodes connected by different hyperedges are the same), thus achieving discriminative pedestrian features.

Influence of the FCGA Module and the HMAL Loss. In this subsection, we evaluate the effectiveness of the FGA and CGA in the FCGA module, as shown in Table V. "w/o FGA" and "w/o CGA" refer to the FCGA module without FGA or CGA, respectively. We also compare models trained with and without HMAL loss, denoted as "+FCGA w/ $\mathcal{L}_{HMAL}$" and "+FCGA w/o $\mathcal{L}_{HMAL}$", to evaluate the influence of the HMAL loss. Besides, we denote the FCGA using the $\mathcal{L}_{HMAL}$ without middle-feature agents and with middle-feature agents as "+FCGA w/ $\mathcal{L}_{HMAL}$ w/o Agents" and "+FCGA w/ $\mathcal{L}_{HMAL}$", respectively.

TABLE V
THE INFLUENCE OF THE FCGA MODULE AND THE HMAL LOSS ON THE SYSU-MM01 AND REGDB DATASETS. R-1 (%) AND MAP (%) ARE REPORTED.

Settings		SYSU-MM01 R-1 / mAP	RegDB R-1 / mAP
Baseline+SLE+HSL		73.3 / 72.4	92.0 / 87.1
+FCGA w/o FGA	w/o $\mathcal{L}_{HMAL}$	73.6 / 72.5	92.4 / 87.6
+FCGA w/o CGA	w/o $\mathcal{L}_{HMAL}$	73.5 / 72.4	92.1 / 87.3
+FCGA	w/o $\mathcal{L}_{HMAL}$	74.1 / 72.7	92.5 / 88.0
+FCGA w/o FGA	w/ $\mathcal{L}_{HMAL}$ w/o Agents	74.1 / 72.9	93.2 / 89.1
+FCGA w/o CGA	w/ $\mathcal{L}_{HMAL}$ w/o Agents	74.5 / 73.4	93.4 / 89.2
+FCGA	w/ $\mathcal{L}_{HMAL}$ w/o Agents	74.6 / 73.9	93.7 / 89.8
+FCGA w/o FGA	w/ $\mathcal{L}_{HMAL}$	74.7 / 73.2	93.9 / 89.6
+FCGA w/o CGA	w/ $\mathcal{L}_{HMAL}$	74.5 / 73.3	94.5 / 90.0
+FCGA	w/ $\mathcal{L}_{HMAL}$	**75.5 / 73.9**	**94.5 / 90.2**

As shown in Table V, on the SYSU-MM01 dataset, the Rank-1 of "+FCGA w/o FCGA w/o $\mathcal{L}_{HMAL}$" and "+FGA w/o CGA w/o $\mathcal{L}_{HMAL}$" methods are 0.3% and 0.2% higher, respectively, than the "Baseline+SLE+HSL" method. When we combine the FGA and the CGA without $\mathcal{L}_{HMAL}$ ("+FCGA w/o $\mathcal{L}_{HMAL}$"), it achieves 74.1% in Rank-1 on the SYSU-MM01 dataset, compared with the "Baseline+SLE+HSL" method. This shows the positive influence of the FCGA module that aligns features at the node-level and region-level perspectives simultaneously to achieve reliable middle features and thus mitigate the modality gap. We adopt the HMAL loss to hierarchically reduce the modality discrepancy with the agents and exchange detailed and semantic information between low- and high-stage networks (i.e., "+FCGA w/ $\mathcal{L}_{HMAL}$") and achieve 73.9% in mAP on the SYSU-MM01 dataset. Furthermore, "+FCGA w/ $\mathcal{L}_{HMAL}$" achieves 0.8% higher Rank-1 than the method without middle-feature agents (i.e., "+FCGA w/ $\mathcal{L}_{HMAL}$ w/o Agents") on the RegDB dataset. These results prove that middle-feature agents and the MDR loss can effectively and hierarchically reduce the modality gap.

TABLE VI
THE INFLUENCE OF THE FEATURE ENHANCEMENT IN THE HMAL LOSS ON THE SYSU-MM01 AND REGDB DATASETS. R-1 (%) AND MAP (%) ARE REPORTED.

Settings	SYSU-MM01 R-1 / mAP	RegDB R-1 / mAP
Baseline+SLE+HSL+FCGA	74.1 / 72.7	92.5 / 88.0
+$\mathcal{L}_{HMAL}$ w/o Low $\leftarrow$ High	74.9 / 73.3	94.0 / 89.5
+$\mathcal{L}_{HMAL}$ w/o Low $\rightarrow$ High	75.2 / 73.5	93.9 / 89.4
+$\mathcal{L}_{HMAL}$	**75.5 / 73.9**	**94.5 / 90.2**

Influence of the Feature Enhancement Loss. Table VI shows that the influence of adopting different feature enhancement strategies (i.e., one-way feature enhancement ("+$\mathcal{L}_{HMAL}$ w/o low $\leftarrow$ high", $\mathcal{L}_{HMAL}$ w/o Low $\rightarrow$ High, +$\mathcal{L}_{HMAL}$ w/o Low $\leftarrow$ High", and "+$\mathcal{L}_{HMAL}$ w/o low $\rightarrow$ high"), and bi-directional feature enhancement ("+$\mathcal{L}_{HMAL}$" and "+$\mathcal{L}_{HMAL}$")). As shown in Table VI, compared to the one-way feature enhancement strategy, the BFE loss ("+$\mathcal{L}_{HMAL}$") can

TABLE VII
THE INFLUENCE OF THE HMAL LOSS AT DIFFERENT STAGES OF THE BACKBONE NETWORK ON THE SYSU-MM01 AND REGDB DATASETS. R-1 (%) AND MAP (%) ARE REPORTED.

Settings	SYSU-MM01	RegDB
	R-1 / mAP	R-1 / mAP
Baseline+SLE+HSL+FCGA	74.1 / 72.7	92.5 / 88.0
+$\mathcal{L}_{HMAL}$ (Stages 1-2)	74.7 / 72.9	94.1 / 90.1
+$\mathcal{L}_{HMAL}$ (Stages 2-3)	75.2 / 73.7	94.2 / 89.9
+$\mathcal{L}_{HMAL}$ (Stages 3-4)	74.8 / 73.1	93.6 / 89.5
+$\mathcal{L}_{HMAL}$ (Stages 1-4)	**75.5 / 73.9**	**94.5 / 90.2**

bring about 0.4% and 0.7% improvements in mAP on the SYSU-MM01 and RegDB datasets, respectively. Notably, the "+$\mathcal{L}_{HMAL}$" surpasses the "Baseline+SLE+HSL+FCGA" by 2.2% in mAP on the RegDB datasets. These results demonstrate that our HMAL loss can better improve the ability of the network to capture discriminative features by performing the bi-directional feature enhancement.

The Influence of the HMAL Loss at Different Stages. We analyze the influence of the different stages of the backbone network for the HMAL loss on the SYSU-MM01 and RegDB datasets to verify its effectiveness. As shown in Table VII, we can observe that the HAML loss achieves the best 75.5% Rank-1 when applied to all the weight-shared feature extraction network stages (i.e., Stages 1-4 of the backbone). This indicates that the HMAL loss can more effectively reduce the modality gap and obtain richer semantic and more detailed feature information when all the stages are equipped.

TABLE VIII
THE INFLUENCE OF EACH TERM IN THE MRIC LOSS ON THE SYSU-MM01 AND REGDB DATASETS. R-1 (%) AND MAP (%) ARE REPORTED.

Settings	SYSU-MM01	RegDB
	R-1 / mAP	R-1 / mAP
Baseline+SLE+HSL+FCGA+$\mathcal{L}_{HMAL}$	75.5 / 73.9	94.5 / 90.2
+$\mathcal{L}_{MRIC}$ w/o $\mathcal{L}_{MRIC}^{MID}$ w/o $\mathcal{L}_{MRIC}^{VIM}$	75.6 / 74.0	94.8 / 90.3
+$\mathcal{L}_{MRIC}$ w/o $\mathcal{L}_{MRIC}^{SL}$ w/o $\mathcal{L}_{MRIC}^{VIM}$	75.5 / 74.3	94.7 / 90.2
+$\mathcal{L}_{MRIC}$ w/o $\mathcal{L}_{MRIC}^{VIM}$	75.7 / 74.2	94.9 / 90.6
+$\mathcal{L}_{MRIC}$	**76.2 / 74.5**	**95.1 / 90.7**

Influence of Each Term in the MRIC Loss. To enhance feature representation and align VIS, IR, and middle features, we propose the MRIC loss ($\mathcal{L}_{MRIC}$), composed of three terms: $\mathcal{L}_{MRIC}^{SL}$, $\mathcal{L}_{MRIC}^{MID}$, and $\mathcal{L}_{MRIC}^{VIM}$. As shown in Table VIII, ablation studies on the RegDB dataset confirm the effectiveness of each term, with mAP progressively increasing from 73.9% to 74.5% as more components are included. Compared with the model trained without MRIC (i.e., "Baseline+SLE+HSL+FCGA+$\mathcal{L}_{HMAL}$"), adding the complete MRIC loss yields a 0.7% Rank-1 improvement on SYSU-MM01 dataset, which demonstrates that the HOH-Net trained with the MRIC loss can obtain a more discriminative and more reliable common feature space.

Influence of the Stage Locations of the Key Components. Our key components (SLE, HSL, and FCGA) can be easily integrated at any stage of the backbone network. In this paper, we utilize the AGW as the backbone for the imaged-based VI-ReID task, which is structured into five stages (i.e., Stages 0-4). We plug these key components into different stages of the backbone and evaluate their performance in Table IX. As shown in Table IX, inserting these components after Stage 3 yields the best performance. This indicates that high-level semantic features from Stage 3 assist the network in better capturing high-order structural information across modalities and ranges, leading to more reliable intermediate features.

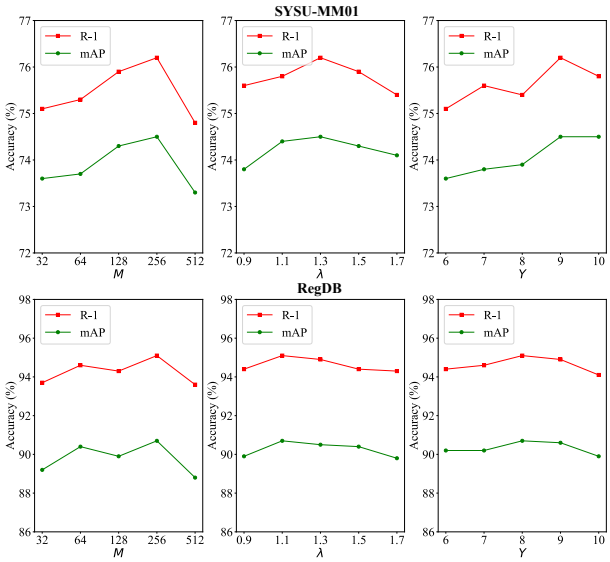


Fig. 6. The influence of the number M of hyperedges, the balancing parameter λ in Eq. (5), and the number Y of person regions in the FCGA module on the SYSU-MM01 and RegDB datasets. R-1 (%) and mAP (%) are reported.

TABLE IX
THE INFLUENCE OF THE BACKBONE STAGE LOCATIONS OF THE KEY COMPONENTS ON THE SYSU-MM01 AND REGDB DATASETS. R-1 (%) AND MAP (%) ARE REPORTED.

Settings	SYSU-MM01	RegDB
	R-1 / mAP	R-1 / mAP
Baseline	69.9 / 66.9	85.0 / 79.1
plugged after Stage 0	71.8 / 68.3	93.1 / 88.4
plugged after Stage 1	70.7 / 67.2	93.8 / 89.2
plugged after Stage 2	72.3 / 69.8	94.1 / 89.5
plugged after Stage 3	**76.2 / 74.5**	**95.1 / 90.7**
plugged after Stage 4	74.2 / 72.0	93.5 / 88.9

TABLE X
THE INFLUENCE OF THE NUMBER OF FRAMES ON THE HITSZ-VCM DATASET. R-1 (%) AND MAP (%) ARE REPORTED.

Frames	HITSZ-VCM	
	VIS to IR	IR to VIS
	R-1 / mAP	R-1 / mAP
12	72.9 / 54.4	70.7 / 53.4
13	73.5 / 56.1	**72.0** / 54.5
14	**74.8 / 57.1**	71.4 / **54.9**
15	73.1 / 55.7	71.2 / 53.9
16	72.6 / 55.2	70.6 / 53.5

Influence of the Hyperparameters. We evaluate the influence of the number M of hyperedges, the hyperparameter λ in

TABLE XI
THE INFLUENCE OF THE DIFFERENT REDUCTION RATIO VALUES ON THE SYSU-MM01 AND REGDB DATASETS. R-1 (%), MAP (%), FLOPS (G), AND PARAMS (M) ARE REPORTED.

Settings	SYSU-MM01	RegDB	FLOPs	Params
	R-1 / mAP	R-1 / mAP		
HOS-Net [12]	75.6 / 74.2	94.7 / 90.4	14.3	83.4
HOH-Net (r=8)	76.1 / **74.6**	94.6 / 90.5	13.6	61.8
HOH-Net (r=16)	75.7 / 74.2	94.9 / 90.6	12.2	59.8
HOH-Net (r=32)	**76.2** / 74.5	**95.1** / **90.7**	11.5	58.8
HOH-Net (r=64)	74.8 / 73.6	94.2 / 89.9	11.1	58.3

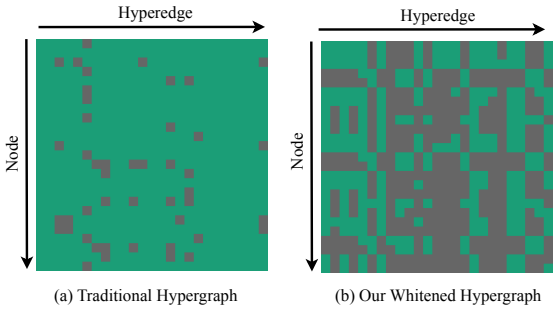
Eq. (5) and the number Y of person regions in the FCGA module in the ranges of {32, 64, 128, 256, 512}, {0.9, 1.1, 1.3, 1.5, 1.7} and {6, 7, 8, 9, 10}, respectively. The results are illustrated in Fig. 6. We first discuss the impact of parameter M on the performance of the proposed HOH-Net. As shown in Fig. 6, the best choice of M is 256 for our HOH-Net on the SYSU-MM01 and RegDB datasets. This shows that an appropriate number of hyperedges can effectively model high-order structure information. Then, we discuss the influence of the parameter λ on model performance by reducing the nodes with low similarity. From Fig. 6, we can observe that the best performance is achieved when the values of λ are set to 1.3 and 1.1 on the SYSU-MM01 and RegDB datasets, respectively. We also change the number Y of person regions from 6 to 10 to explore the most effective setting of the proposed HOH-Net. As shown in Fig. 6, the best results of the HOH-Net are achieved when Y = 9 and Y = 8 on the SYSU-MM01 and RegDB datasets, respectively.

We also analyze the impact of different video frames on video-based VI-ReID, and the corresponding results are shown in Table X. From Table X, we can observe that the best mAP performance is achieved when using 14 frames. This indicates that our HOH-Net is capable of extracting discriminative pedestrian features according to the temporal information, and effectively reduces the modality gap by reliable middle-feature feature agents, thereby improving the performance of cross-modality retrieval in video-based scenarios.

Besides, we introduce the squeeze and excitation by using the reduction ratio r to reduce the number of trainable parameters in the HOH-Net. As shown in Table XI, when we set the reduction ratio r to 32, the proposed HOH-Net achieves 76.2% and 95.1% in Rank-1 on these two datasets, respectively, and compared to the HOS-Net [12], the number of parameters of the HOH-Net is reduced by 29.5%, while our method can efficiently generate middle features.

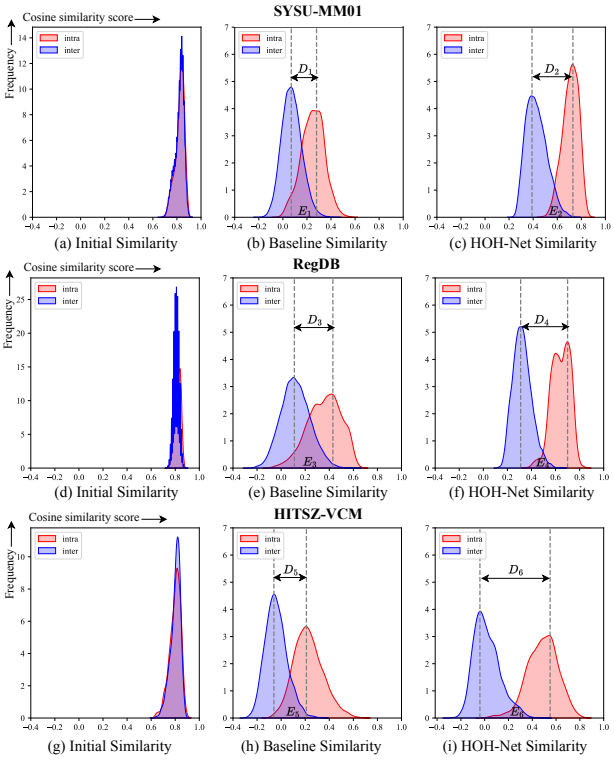
D. Visualization Analysis

Visualization of High-Order Relationship. As illustrated in Fig. 7, we present visualizations of the high-order relationships derived from the traditional hypergraph network and our whitened hypergraph network on the SYSU-MM01 dataset. For the traditional hypergraph network, many nodes share the same hyperedges, and thus the diverse and complex high-order connections collapse into a single connection. In contrast, our whitened hypergraph network prevents model collapse by using a whitening operation to project the feature nodes into a spherical distribution.



(a) Traditional Hypergraph (b) Our Whitened Hypergraph

Fig. 7. Visualization of the high-order relationship obtained by different methods. In each column, the green square represents that the node is connected by a hyperedge, while the gray square represents that the node has no connections.



(a) Initial Similarity (b) Baseline Similarity (c) HOH-Net Similarity
(d) Initial Similarity (e) Baseline Similarity (f) HOH-Net Similarity
(g) Initial Similarity (h) Baseline Similarity (i) HOH-Net Similarity

Fig. 8. (a-i) give the distributions of the intra-class and inter-class similarities of cross-modality features from the initial network, the Baseline, and our HOH-Net on the SYSU-MM01, RegDB, and HITSZ-VCM datasets, respectively. Intra-class and inter-class represent the positive (for the same identity) and the negative (for the different identities) matching from the VIS and IR modalities, respectively. The larger the distance D and the smaller the overlapping area E, the better the model performance.

Feature Distribution Visualization. We randomly select 90,000 positive and negative pairs from the query and gallery sets and visualize the cosine similarity distributions on SYSU-MM01, RegDB, and HITSZ-VCM (see Figs. 8(a–i)). As shown in Figs. 8(b–c), (e–f), and (h-i), the HOH-Net shows larger intra-/inter-class distribution differences (D_2, D_4, and

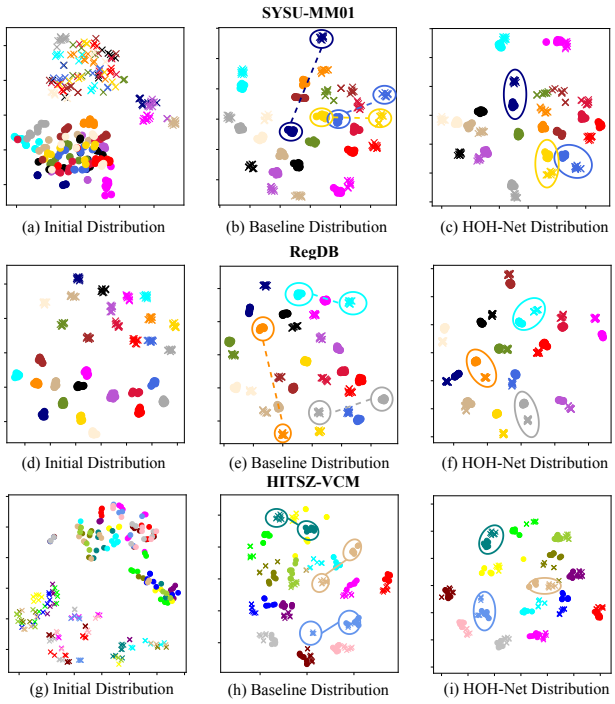


Fig. 9. (a)/(d)/(g), (b)/(e)/(h), and (c)/(f)/(i) visualize the distributions of sample features from the initial network, the Baseline, and our HOH-Net by t-SNE [61] on the SYSU-MM01, RegDB, and HITSZ-VCM datasets, respectively. Circles and crosses represent features from VIS and IR modalities, respectively. Different colors represent different people. A total of 15 persons are chosen from the query and gallery sets of the SYSU-MM01, RegDB, and HITSZ-VCM datasets, respectively.

Fig. 10. Attention maps and retrieval results obtained by the Baseline and the proposed HOH-Net on the SYSU-MM01 and HITSZ-VCM datasets. (a) Query images. (b) Attention maps. (c) Retrieval results (Green: correct retrieval, Red: incorrect retrieval).

D_6) and smaller overlap areas (E_2, E_4, and E_6), compared to Baseline (D_1, D_3, D_5, E_1, E_3, and E_5), respectively. This demonstrates that the proposed HOH-Net effectively reduces the modality gap between different modalities. We also use t-SNE [61] to illustrate the feature distributions obtained by different retrieval results produced by both our method and the Baseline on the SYSU-MM01, RegDB, and HITSZ-VCM datasets, as depicted in Figs. 9(a-i). Compared with the Baseline (see Figs. 9(b), (e), and (h)), the distance between features of the same pedestrian obtained by our method (see Figs. 9(c), (f), and (i)) is more compact on the SYSU-MM01, RegDB, and HITSZ-VCM datasets. This phenomenon demonstrates that our HOH-Net is able to achieve a more discriminative common feature space for effective VI-ReID, and the discrepancy between VIS and IR modalities can be effectively mitigated.

Retrieval Results. To further evaluate the effectiveness of our proposed HOH-Net, we provide the attention maps and retrieval results produced by both the Baseline and our method on the SYSU-MM01 and HITSZ-VCM datasets, respectively. As illustrated in Fig. 10(b), different from the Baseline, our HOH-Net exhibits a superior ability to focus on discriminative features, reducing the impact of pose variations, background interference, and modality gap. From Fig. 10(c), we can observe that our HOH-Net can perform more accurate cross-modality pedestrian retrieval, while the Baseline is easily affected by modality discrepancy, resulting in poor retrieval results.

V. Conclusion

In this paper, we propose a novel HOH-Net mainly consisting of the HSL module, and the FCGA module with the HMAL loss and the MRIC loss for VI-ReID. The HSL module exploits diverse and complex high-order structure information of shot- and long-range features that are extracted from the SLE module and prevents model collapse by employing a whitened hypergraph. Moreover, the FCGA module generates reliable middle features from the node-level and region-level perspectives. In particular, the HMAL loss hierarchically reduces the modality gap by leveraging the middle-feature agents and performs the bi-directional feature enhancement between different stages to obtain the discriminative features. Finally, the MRIC loss minimizes the distance between the VIS, IR, and middle features, thereby establishing a discriminative and reliable common feature space. The quantitative and qualitative experiments on the four challenging VI-ReID datasets confirm the superiority of the HOH-Net in comparison with several state-of-the-art methods. Our HOH-Net achieves

impressive performance, but one limitation of our method could be the effective learning with limited data, which is also a common disadvantage of VI-ReID methods. In future work, we will focus on few-shot or zero-shot learning and further improve the generalization performance of the model.

REFERENCES

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

[2] M. Ye, S. Chen, C. Li, W.-S. Zheng, D. Crandall, and B. Du, "Transformer for object re-identification: A survey," *Int. J. Comput. Vis.*, vol. 133, pp. 2410–2440, 2025.

[3] G. Zhang, Y. Yang, Y. Zheng, G. Martin, and R. Wang, "Mask-aware hierarchical aggregation transformer for occluded person re-identification," *IEEE Trans. Circuit Syst. Video Technol.*, pp. 1–1, 2025.

[4] Z. Wei, X. Yang, N. Wang, and X. Gao, "Rbdf: Reciprocal bidirectional framework for visible infrared person reidentification," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10 988–10 998, 2022.

[5] J. Liu, J. Wang, N. Huang, Q. Zhang, and J. Han, "Revisiting modality-specific feature compensation for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7226–7240, 2022.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[7] H. Liu, X. Tan, and X. Zhou, "Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification," *IEEE Trans. Multimedia*, vol. 23, pp. 4414–4425, 2020.

[8] H. Lu, X. Zou, and P. Zhang, "Learning progressive modality-shared Transformers for effective visible-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 1835–1843.

[9] D. Zhang, Z. Zhang, Y. Ju, C. Wang, Y. Xie, and Y. Qu, "Dual mutual learning for cross-modality person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5361–5373, 2022.

[10] J. Zhao, H. Wang, Y. Zhou, R. Yao, S. Chen, and A. El Saddik, "Spatial-channel enhanced Transformer for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 3668–3680, 2023.

[11] T. Liang, Y. Jin, W. Liu, and Y. Li, "Cross-modality transformer with modality mining for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 8432–8444, 2023.

[12] 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.*, 2024, pp. 4596–4604.

[13] G. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z. Hou, "Cross-modality paired-images generation for rgb-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12 144–12 151.

[14] H. Li, M. Liu, Z. Hu, F. Nie, and Z. Yu, "Intermediary-guided bidirectional spatial–temporal aggregation network for video-based visible-infrared person re-identification," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 33, no. 9, pp. 4962–4972, 2023.

[15] G. Du and L. Zhang, "Enhanced invariant feature joint learning via modality-invariant neighbor relations for cross-modality person re-identification," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 4, pp. 2361–2373, 2024.

[16] R. Zhang, Z. Cao, Y. Huang, S. Yang, L. Xu, and M. Xu, "Visible-infrared person re-identification with real-world label noise," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 35, no. 5, pp. 4857–4869, 2025.

[17] X. Ni, L. Yuan, and K. Lv, "Efficient single-object tracker based on local-global feature fusion," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 2, pp. 1114–1122, 2024.

[18] S. Chen, H. Da, D.-H. Wang, X.-Y. Zhang, Y. Yan, and S. Zhu, "Hasi: Hierarchical attention-aware spatio–temporal interaction for video-based person re-identification," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 6, pp. 4973–4988, 2024.

[19] L. Guo, L. Jin, and E. Song, "Queue-augmented correlation-biased orthogonality loss and implicit selective transformer for facial expression recognition in the wild," *IEEE Trans. Circuit Syst. Video Technol.*, pp. 1–1, 2025.

[20] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, "Structure-aware positional Transformer for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 2352–2364, 2022.

[21] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.

[22] A. Wu, W. Zheng, S. Gong, and J. Lai, "Rgb-ir person re-identification by cross-modality similarity preservation," *Int. J. Comput. Vis.*, vol. 128, pp. 1765–1785, 2020.

[23] X. Cheng, H. Yu, K. H. M. Cheng, Z. Yu, and G. Zhao, "Mdanet: Modality-aware domain alignment network for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 27, pp. 2015–2027, 2025.

[24] M. Liu, Z. Zhang, Y. Bian, X. Wang, Y. Sun, B. Zhang, and Y. Wang, "Cross-modality semantic consistency learning for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 27, pp. 568–580, 2025.

[25] Y. Zhang and H. Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2153–2162.

[26] K. Jiang, T. Zhang, X. Liu, B. Qian, Y. Zhang, and F. Wu, "Cross-modality transformer for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 480–496.

[27] H. Li, M. Li, Q. Peng, S. Wang, H. Yu, and Z. Wang, "Correlation-guided semantic consistency network for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2361–2373, 2024.

[28] X. Yang, W. Dong, M. Li, Z. Wei, N. Wang, and X. Gao, "Cooperative separation of modality shared-specific features for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 26, pp. 8172–8183, 2024.

[29] X. Tian, Z. Zhang, C. Wang, W. Zhang, Y. Qu, L. Ma, Z. Wu, Y. Xie, and D. Tao, "Variational distillation for multi-view learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 4551–4566, 2024.

[30] Y. Zhang, Y. Liang, J. Leng, and Z. Wang, "Scgtracker: Spatio-temporal correlation and graph neural networks for multiple object tracking," *Pattern Recognition*, vol. 149, 2024, Art. no. 110249.

[31] X. Zhang, K. Liu, X. Wang, Z. Zhou, and H. Chen, "Rmgnet: The progressive relationship-mining graph neural network for text-to-image person re-identification," *IEEE Trans. Circuit Syst. Video Technol.*, pp. 1–1, 2025.

[32] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 3558–3565.

[33] G. Wadhwa, A. Dhall, S. Murala, and U. Tariq, "Hyperrealistic image inpainting with hypergraphs," in *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3912–3921.

[34] Y. Han, P. Wang, S. Kundu, Y. Ding, and Z. Wang, "Vision hgnn: An image is more than a graph of nodes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 19 878–19 888.

[35] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 1559–1572, 2022.

[36] X. Yang, C. Deng, T. Liu, and D. Tao, "Heterogeneous graph attention network for unsupervised multiple-target domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1992–2003, 2022.

[37] D. Guo, Y. Shao, Y. Cui, Z. Wang, L. Zhang, and C. Shen, "Graph attention tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9543–9552.

[38] D. J. Higham and H.-L. de Kergorlay, "Mean field analysis of hypergraph contagion models," *SIAM. J. Appl. Math.*, vol. 82, no. 6, pp. 1987–2007, 2022.

[39] H. Park, S. Lee, J. Lee, and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12 046–12 055.

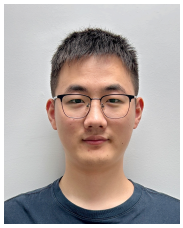
[40] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1092–1099.

[41] R. Rubinstein, "The cross-entropy method for combinatorial and continuous optimization," *Methodol. Comput. Appl.*, vol. 1, pp. 127–190, 1999.

[42] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[43] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors.*, vol. 17, no. 3, p. 605, 2017.

[44] X. Lin, J. Li, Z. Ma, H. Li, S. Li, K. Xu, G. Lu, and D. Zhang, "Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20 973–20 982.
[45] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13 567–13 576.
[46] J. Wu, H. Liu, Y. Su, W. Shi, and H. Tang, "Learning concordant attention via target-aware alignment for visible-infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11 122–11 131.
[47] M. Ye, Z. Wu, C. Chen, and B. Du, "Channel augmentation for visible-infrared re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, 2024.
[48] S. Chan, W. Meng, C. Bai, J. Hu, and S. Chen, "Diverse-feature collaborative progressive learning for visible-infrared person re-identification," *IEEE Trans Ind. Informat.*, vol. 20, no. 5, pp. 7754–7763, 2024.
[49] Z. Lu, R. Lin, and H. Hu, "Disentangling modality and posture factors: Memory-attention and orthogonal decomposition for visible-infrared person re-identification," *IEEE Trans. Neural Networks Learn. Syst.*, 2024.
[50] M. Alehdaghi, A. Josi, R. M. O. Cruz, P. Shamsolmoali, and E. Granger, "Adaptive generation of privileged intermediate information for visible-infrared person re-identification," *IEEE Trans. Inf. Forensics Secur.*, vol. 20, pp. 3400–3413, 2025.
[51] Y. Du, C. Lei, Z. Zhao, Y. Dong, and F. Su, "Video-based visible-infrared person re-identification with auxiliary samples," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 1313–1325, 2024.
[52] W. Hou, W. Wang, Y. Yan, D. Wu, and Q. Xia, "A three-stage framework for video-based visible-infrared person re-identification," *IEEE Signal Process Lett.*, vol. 31, pp. 1254–1258, 2024.
[53] C. Zhou, J. Li, H. Li, G. Lu, Y. Xu, and M. Zhang, "Video-based visible-infrared person re-identification via style disturbance defense and dual interaction," in *Proceedings of the ACM Int. Conf. Multimedia*, 2023, pp. 46–55.
[54] Y. Feng, F. Chen, J. Yu, Y. Ji, F. Wu, T. Liu, S. Liu, X.-Y. Jing, and J. Luo, "Cross-modality spatial-temporal transformer for video-based visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 26, pp. 6582–6594, 2024.
[55] Z. Wei, X. Yang, N. Wang, and X. Gao, "Dual-adversarial representation disentanglement for visible infrared person re-identification," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 2186–2200, 2024.
[56] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: chasing higher flops for faster neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12 021–12 031.
[57] Z. Chen, Z. He, and Z.-M. Lu, "Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention," *IEEE Trans. Image Process.*, 2024.
[58] Y. Duan, W. Wang, Z. Chen, X. Zhu, L. Lu, T. Lu, Y. Qiao, H. Li, J. Dai, and W. Wang, "Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures," *Proc. Int. Conf. Learn. Represent.*, 2025.
[59] K. Han, Y. Wang, J. Guo, Y. Tang, and E. Wu, "Vision gnn: An image is worth graph of nodes," *Adv. Neural Inform. Process. Syst.*, vol. 35, pp. 8291–8303, 2022.
[60] A. Hatamizadeh and J. Kautz, "Mambavision: A hybrid mamba-transformer vision backbone," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 25 261–25 270.
[61] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.



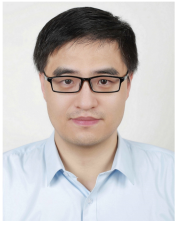
Liuxiang Qiu received the M.S. degree from Xiamen University of Technology, China, in 2024. His current research interests include computer vision, pattern recognition, and person re-identification.

Si Chen (Senior Member, IEEE) received the Ph.D. degree from the School of Informatics, Xiamen University, China, in 2014. She is currently a Full Professor with the School of Computer and Information Engineering, Xiamen University of Technology, China. In recent years, she has published more than 40 papers in international journals and conferences, including International Journal of Computer Vision, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, Pattern Recognition, CVPR, AAAI, and ACM MM. Her research interests include computer vision, machine learning, and pattern recognition.

Jing-Hao Xue (Senior Member, IEEE) received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998, and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is currently a Full Professor with the Department of Statistical Science, University College London. His research interests include statistical pattern recognition, machine learning, and computer vision. He received the Best Associate Editor Award of 2021 from the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the Outstanding Associate Editor Award of 2022 from the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Da-Han Wang (Member, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, China, in 2012. He was a Post-Doctoral Fellow with the School of Informatics, Xiamen University, from 2012 to 2014. He is currently a Full Professor and the Vice Dean with the School of Computer and Information Engineering, Xiamen University of Technology. He is also the Director of the Fujian Key Laboratory of Pattern Recognition and Image Understanding. His research interests include pattern recognition, text detection and recognition, and semantic segmentation.

Shunzhi Zhu received the Ph.D. degree from the School of Informatics, Xiamen University, China, in 2007. He was a Visiting Professor with Florida International University, USA, from 2013 to 2014. He is currently a Full Professor and the Vice-Principal of Xiamen University of Technology. His research interests include data mining, video analysis and processing, information recommendation, and system engineering.



Yan Yan (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Tsinghua University, China, in 2009. He was a Research Engineer with Nokia Japan Research and Development Center from 2009 to 2010 and a Project Leader with Panasonic Singapore Laboratory in 2011. He is currently a Full Professor with the School of Informatics, Xiamen University, China. He has published more than 100 papers in international journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, International Journal of Computer Vision, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, Pattern Recognition, ICCV, CVPR, ECCV, ACM MM, and AAAI. His research interests include computer vision and pattern recognition.