High-Order Structure Based Middle-Feature Learning for Visible-Infrared Person Re-identification

Liuxiang Qiu[1,2*], **Si Chen**[2*], **Yan Yan**[1†], **Jing-Hao Xue**[3], **Da-Han Wang**[2], **Shunzhi Zhu**[2]

[1]Xiamen University, China
[2] Xiamen University of Technology, China
[3] University College London, UK

Abstract

Visible-infrared person re-identification (VI-ReID) aims to retrieve images of the same persons captured by visible (VIS) and infrared (IR) cameras. Existing VI-ReID methods ignore high-order structure information of features while being relatively difficult to learn a reasonable common feature space due to the large modality discrepancy between VIS and IR images. To address the above problems, we propose a novel high-order structure based middle-feature learning network (HOS-Net) for effective VI-ReID. Specifically, we first leverage a short- and long-range feature extraction (SLE) module to effectively exploit both short-range and long-range features. Then, we propose a high-order structure learning (HSL) module to successfully model the high-order relationship across different local features of each person image based on a whitened hypergraph network. This greatly alleviates model collapse and enhances feature representations. Finally, we develop a common feature space learning (CFL) module to learn a discriminative and reasonable common feature space based on middle features generated by aligning features from different modalities and ranges. In particular, a modality-range identity-center contrastive (MRIC) loss is proposed to reduce the distances between the VIS, IR, and middle features, smoothing the training process. Extensive experiments on the SYSU-MM01, RegDB, and LLCM datasets show that our HOS-Net achieves superior state-of-the-art performance. Our code is available at https://github.com/Jaulaucoeng/HOS-Net.

Introduction

Over the past few years, person re-identification (ReID) has attracted increasing attention due to its significant importance in surveillance and security applications. A large number of single-modality person ReID methods have been proposed based on visible (VIS) cameras. However, these methods may fail under low-light conditions. Unlike VIS cameras, infrared (IR) cameras are less affected by illumination changes. Recently, visible-infrared person re-identification (VI-ReID), which leverages both VIS and IR cameras, has been developed to match cross-modality person images, mitigating the limitations of single-modality person ReID.

A major challenge of VI-ReID is the large modality discrepancy between VIS and IR images. To reduce the modality discrepancy, existing VI-ReID methods can be divided into image-level and feature-level methods. The image-level methods (Dai et al. 2018; Wang et al. 2020; Wei et al. 2022), generate middle-modality or new modality images based on generative adversarial networks (GAN). However, the GAN-based methods easily suffer from the problems of color inconsistency or loss of image details. Hence, the generated images may not be reliable for subsequent classification.

The feature-level methods (Ye et al. 2021b; Lu, Zou, and Zhang 2023; Zhang et al. 2022) follow a two-step learning pipeline (i.e., they first extract features for VIS and IR images, and then map these features into a common feature space). Generally, these methods have two issues. On the one hand, they often ignore high-order structure information of features (i.e., the different levels of dependence across local features), which can be important for matching cross-modality images. On the other hand, they usually directly minimize the distances between VIS and IR features in the common feature space. However, such a manner increases the difficulty of learning a reasonable common feature space due to the large modality discrepancy.

To address the above issues, in this paper, we propose a novel high-order structure based middle-feature learning network (HOS-Net), which consists of a backbone, a short- and long-range feature extraction (SLE) module, a high-order structure learning (HSL) module, and a common feature space learning (CFL) module, for VI-ReID. The key novelty of our method lies in the novel formulation of exploiting *high-order structure information* and *middle features* to learn a discriminative and reasonable common feature space, greatly alleviating the modality discrepancy.

Specifically, given a VIS-IR image pair, the SLE module (consisting of a convolutional branch and a Transformer branch) extracts short-range and long-range features. Then, the HSL module models the dependence on short-range and long-range features based on a whitened hypergraph. Finally, the CFL module learns a common feature space by generating and leveraging middle features. In the CFL module, instead of directly adding or concatenating features from different modalities and ranges, we leverage graph attention to properly align these features, obtaining middle features. Based on it, a modality-range identity-center contrastive

[*]These authors contributed equally.

[†]Corresponding author (email: yanyan@xmu.edu.cn).

(MRIC) loss is developed to reduce the distances between the VIS, IR, and middle features, smoothing the process of learning the common feature space.

The contributions of our work are twofold:

- First, we introduce an HSL module to learn high-order structure information of both short-range and long-range features. Such an innovative way effectively models high-order relationship across different local features of a person image without suffering from model collapse, greatly enhancing feature representations.
- Second, we design a CFL module to learn a discriminative and reasonable common feature space by taking advantage of middle features. In particular, a novel MRIC loss is developed to minimize the distances between VIS, IR, and middle features. This is beneficial for the extraction of discriminative modality-irrelevant ReID features.

Extensive experiments on the SYSU-MM01, RegDB, and LLCM datasets demonstrate that our proposed HOS-Net obtains excellent performance in comparison with several state-of-the-art VI-ReID methods. The full version of this paper, including supplement, can be found at https://arxiv.org/abs/2312.07853.

Related Work

Single-Modality Person Re-Identification (ReID). A variety of single-modality person ReID methods have been developed and achieved promising performance in the cases of occlusion, cloth-changing, and pose changes. Yan *et al.* (Yan et al. 2021) propose an occlusion-based data augmentation strategy and a bounded exponential distance loss for occluded person ReID. Jin *et al.* (Jin et al. 2022) introduce an additional gait recognition task to learn cloth-agnostic features. Note that these methods are based on VIS cameras and thus they perform poorly in low-light conditions.

Visible-Infrared Person Re-Identification (VI-ReID). The image-level methods (Dai et al. 2018; Wang et al. 2020; Wei et al. 2022) often reduce the modality discrepancy by generating middle-modality images or new modality images. Wei *et al.* (Wei et al. 2022) propose a bidirectional image translation subnetwork to generate middle-modality images from VIS and IR modalities. Li *et al.* (Li et al. 2020) and Zhang *et al.* (Zhang et al. 2021) introduce light-weight middle-modality image generators to mitigate the modality discrepancy. Instead of generating middle-modality images, we align the features from different modalities and ranges with graph attention, generating reliable middle features. Moreover, we design an MRIC loss to optimize the distances between VIS, IR, and middle features, benefiting the extraction of discriminative ReID features.

The feature-level methods map the features of different modalities into a common feature space to reduce the modality discrepancy. A few methods (Ye et al. 2021b; Yang, Chen, and Ye 2023; Lu, Zou, and Zhang 2023) leverage CNN or ViT as the backbone to extract features. Some methods (Chen et al. 2022a; Wan et al. 2023) adopt off-the-shelf key point extractors to generate key point labels of person images and learn modality-irrelevant features. But the key point extractor may introduce noisy labels, deteriorating the discriminability of final ReID features. Many VI-ReID methods (Liu, Tan, and Zhou 2020; Huang et al. 2022, 2023) employ the contrastive-based loss, which directly minimizes the distances between VIS and IR features, to obtain a common feature space. However, it is not a trivial task to learn a reasonable common feature space due to the large modality discrepancy between modalities.

Our method belongs to feature-level methods. However, conventional feature-level methods mainly consider first-order structure information of features (i.e., the pairwise relation across features). Moreover, they directly reduce the distances between VIS and IR features. Different from these methods, our method not only captures high-order structure information of features but also generates middle features, greatly facilitating our model to learn an effective common feature space.

Graph Neural Networks in Person Re-Identification. Graph neural network (GNN) is a class of neural networks that is designed to operate on graph-structured data. Li *et al.* (Li et al. 2021) propose a pose and similarity based-GNN to reduce the problem of pose misalignment for single-modality person ReID. Wan *et al.* (Wan et al. 2023) develop a geometry guided dual-alignment strategy to align VIS and IR features, improving the consistency of multi-modality node representations. Different from pairwise connections in the vanilla graph models, Feng *et al.* (Feng et al. 2019) propose a hypergraph neural network (HGNN) to encode high-order feature correlations in a hypergraph structure. Lu *et al.* (Lu et al. 2023) model high-order spatio-temporal correlations based on HGNN (which relies on high-quality skeleton labels) for video person ReID.

However, the above HGNN-based methods may easily suffer from the model collapse problem (i.e., high-order correlations collapse to a single correlation) since a hyperedge can connect an arbitrary number of nodes. Unlike the above methods, we leverage the whitening operation, which plays the role of "scattering" on the nodes of the hypergraph, to significantly alleviate model collapse.

Proposed Method

Overview

The overview of our proposed HOS-Net is given in Figure 1. HOS-Net consists of a backbone, an SLE module, an HSL module, and a CFL module. In this paper, we adopt a two-stream AGW (Ye et al. 2021b) as the backbone. Given a VIS-IR image pair with the same identity, we first pass it through the backbone to obtain paired VIS-IR features. Then, these features are fed into the SLE module to learn short-range and long-range features for each modality. Next, the HSL module exploits high-order structure information of short-range and long-range features based on a whitened hypergraph network. Finally, the CFL module learns a discriminative common feature space based on middle features that are obtained by aligning VIS and IR features through graph attention. In the CFL module, 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.

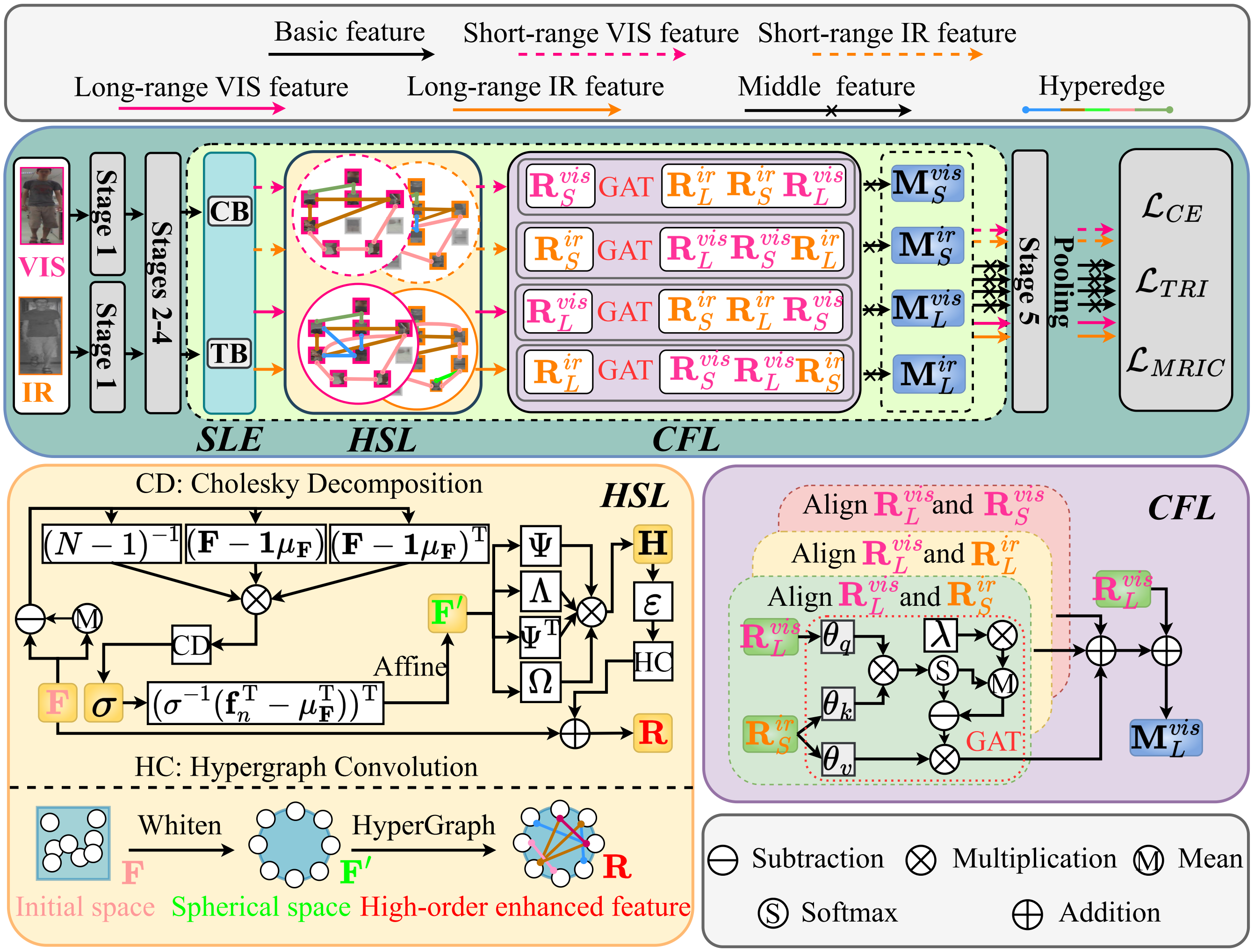
Basic feature
Short-range VIS feature
Short-range IR feature
Long-range VIS feature
Long-range IR feature
Middle feature
Hyperedge
VIS
IR
Stage 1
Stage 1
Stages 2-4
CB
TB
$\mathbf{R}_S^{vis}$ GAT $\mathbf{R}_L^{ir}$ $\mathbf{R}_S^{ir}$ $\mathbf{R}_L^{vis}$
$\mathbf{R}_S^{ir}$ GAT $\mathbf{R}_L^{vis}$ $\mathbf{R}_S^{vis}$ $\mathbf{R}_L^{ir}$
$\mathbf{R}_L^{vis}$ GAT $\mathbf{R}_S^{ir}$ $\mathbf{R}_L^{ir}$ $\mathbf{R}_S^{vis}$
$\mathbf{R}_L^{ir}$ GAT $\mathbf{R}_S^{vis}$ $\mathbf{R}_L^{vis}$ $\mathbf{R}_S^{ir}$
$\mathbf{M}_S^{vis}$
$\mathbf{M}_S^{ir}$
$\mathbf{M}_L^{vis}$
$\mathbf{M}_L^{ir}$
Stage 5
Pooling
$\mathcal{L}_{CE}$
$\mathcal{L}_{TRI}$
$\mathcal{L}_{MRIC}$
SLE
HSL
CFL
CD: Cholesky Decomposition
HSL
$(N-1)^{-1}$
$(\mathbf{F}-\mathbf{1}\mu_\mathbf{F})$
$(\mathbf{F}-\mathbf{1}\mu_\mathbf{F})^\mathrm{T}$
Ψ
Λ
Ψ^T
Ω
$\mathbf{H}$
ε
HC
M
CD
$\mathbf{F}'$
Affine
$\mathbf{F}$
σ
$(\sigma^{-1}(\mathbf{f}_n^\mathrm{T}-\mu_\mathbf{F}^\mathrm{T}))^\mathrm{T}$
$\mathbf{R}$
HC: Hypergraph Convolution
Whiten
HyperGraph
$\mathbf{F}$
$\mathbf{F}'$
$\mathbf{R}$
Initial space
Spherical space
High-order enhanced feature
Align $\mathbf{R}_L^{vis}$ and $\mathbf{R}_S^{vis}$
Align $\mathbf{R}_L^{vis}$ and $\mathbf{R}_L^{ir}$
Align $\mathbf{R}_L^{vis}$ and $\mathbf{R}_S^{ir}$
CFL
$\mathbf{R}_L^{vis}$
θ_q
λ
S
M
θ_k
$\mathbf{R}_S^{ir}$
θ_v
GAT
$\mathbf{R}_L^{vis}$
$\mathbf{M}_L^{vis}$
Subtraction
Multiplication
M Mean
S Softmax
Addition

Figure 1: Overview of the proposed HOS-Net, including a backbone, a short- and long-range feature extraction (SLE) module, a high-order structure learning (HSL) module, and a common feature space learning (CFL) module. The HOS-Net is jointly optimized by $\mathcal{L}_{CE}$, $\mathcal{L}_{TRI}$, and $\mathcal{L}_{MRIC}$.

Short- and Long-Range Feature Extraction (SLE) Module

Conventional VI-ReID methods (Ye et al. 2021b; Yang, Chen, and Ye 2023) often leverage CNN or ViT for feature extraction. CNN excels at capturing short-range features, while ViT is good at obtaining long-range features (Zhang, Hu, and Wang 2022; Chen et al. 2022b). In this paper, we adopt an SLE module to exploit short-range and long-range features by taking advantage of both CNN and ViT. The SLE module contains a convolutional branch (CB) and a Transformer branch (TB). CB contains 3 convolutional blocks and TB contains 2 Transformer blocks with 4 heads. Assume that we have a VIS-IR image pair $\{\mathbf{I}^{vis}, \mathbf{I}^{ir}\}$ with the same identity. We denote the VIS and IR features obtained from the backbone as $\mathbf{B}^{vis}$ and $\mathbf{B}^{ir}$, respectively.

Then, $\mathbf{B}^{vis}$ and $\mathbf{B}^{ir}$ are fed into the SLE module to obtain short-range and long-range features for each modality, i.e.,

$$\begin{aligned} \mathbf{F}_S^{vis} &= \mathrm{CB}(\mathbf{B}^{vis}), \quad \mathbf{F}_L^{vis} = \mathrm{TB}(\mathbf{B}^{vis}), \\ \mathbf{F}_S^{ir} &= \mathrm{CB}(\mathbf{B}^{ir}), \quad \mathbf{F}_L^{ir} = \mathrm{TB}(\mathbf{B}^{ir}), \end{aligned} \tag{1}$$

where $\mathrm{CB}(\cdot)$ and $\mathrm{TB}(\cdot)$ represent the convolutional branch and the Transformer branch, respectively; $\mathbf{F}_S^{vis} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{F}_S^{ir} \in \mathbb{R}^{H\times W\times C}$ denote the short-range features for the VIS and IR images, respectively; $\mathbf{F}_L^{vis} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{F}_L^{ir} \in \mathbb{R}^{H\times W\times C}$ denote the long-range features for the VIS and IR images, respectively; H, W, and C denote the height, width, and channel number of the feature, respectively. Thus, for a VIS-IR image pair, we can obtain the feature set $\mathcal{Q} = \{\mathbf{F}_L^{vis}, \mathbf{F}_S^{vis}, \mathbf{F}_L^{ir}, \mathbf{F}_S^{ir}\}$, which is used as the input of the HSL module.

High-Order Structure Learning (HSL) Module

The features extracted from the SLE module only encode pixel-wise and region-wise dependencies in the person images. However, the high-order structure information, which indicates different levels of relation in the features (e.g., head, torso, upper arm, and lower arm belongs to the upper part of the body while head, torso, arm, and leg belong to the whole body), is not well exploited. Therefore, inspired by HGNN (Feng et al. 2019), we introduce the HSL module to capture high-order correlations across different local features, enhancing feature representations. Note that the conventional HGNN tends to suffer from the problem of model collapse. To alleviate this problem, we take advantage of the whitening operation and apply it to the hypergraph network, as shown in Figure 1.

Different from pairwise connections in the vanilla graph models, a hypergraph can connect an arbitrary number of nodes to exploit high-order structure information. We construct a whitened hypergraph for each feature in $\mathcal{Q}$. The hypergraph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathcal{V} = \{v_1, \cdots, v_N\}$ denotes the node set, $\mathcal{E} = \{e_1, \cdots, e_M\}$ denotes the hyperedge set, and $\mathbf{W}$ represents the weight matrix of the hyperedge set. Here, $N = HW$ and M are the numbers of nodes and hyperedges, respectively. In this paper, we consider each $1 \times 1 \times C$ grid of each feature in $\mathcal{Q}$ as a node. We represent the n-th node by $\mathbf{f}_n \in \mathbb{R}^{1\times C}$ and thus all nodes are represented by $\mathbf{F} = [\mathbf{f}_1; \cdots; \mathbf{f}_N] \in \mathbb{R}^{N\times C}$.

The traditional hypergraph network allows for unrestricted connections among nodes to capture high-order structure information. Hence, it easily suffers from model collapse (i.e., the nodes connected by different hyperedges are the same) during hypergraph learning. To mitigate this problem, we introduce a whitening operation to project the nodes into a spherical distribution. In fact, the whitening operation plays the role of "scattering" on the nodes, thereby avoiding the collapse of diverse high-order connections to a single connection. This enables us to explore various high-order relationships across these features effectively.

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

$$\mathbf{f}'_n = \gamma_n(\sigma^{-1}(\mathbf{f}_n^{\mathrm{T}} - \mu_{\mathbf{F}}^{\mathrm{T}}))^{\mathrm{T}} + \beta_n, \tag{2}$$

where $\sigma \in \mathbb{R}^{C\times C}$ denotes the lower triangular matrix, which 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 affine parameters learned from the network. In this way, all the whitened nodes can be represented by $\mathbf{F}' = [\mathbf{f}'_1; \cdots; \mathbf{f}'_N] \in \mathbb{R}^{N\times C}$.

Similar to (Higham and de Kergorlay 2022), we use cross-correlation to learn the incidence matrix $\mathbf{H} \in \mathbb{R}^{N\times M}$, i.e.,

$$\mathbf{H} = \varepsilon(\Psi(\mathbf{F}')\Lambda(\mathbf{F}')\Psi(\mathbf{F}')^{\mathrm{T}}\Omega(\mathbf{F}')), \tag{3}$$

where $\Psi(\cdot)$ represents the linear transformation; $\Lambda(\cdot)$ and $\Omega(\cdot)$ are responsible for learning a distance metric by a diagonal operation and determining the contribution of the node to the corresponding hyperedges through the learnable parameters, respectively; $\varepsilon(\cdot)$ is the step function.

Based on the learned $\mathbf{H}$, we adopt a hypergraph convolutional operation to aggregate high-order structure information and then enhance feature representations. The relation enhanced feature $\mathbf{R} \in \mathbb{R}^{N\times C}$ can be obtained as

$$\mathbf{R} = (\mathbf{I} - \mathbf{D}^{1/2}\mathbf{H}\mathbf{W}\mathbf{B}^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{D}^{-1/2})\mathbf{F}'\Theta + \mathbf{F}, \tag{4}$$

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 process, we pass features in $\mathcal{Q}$ through the HSL module and obtain a relation-enhanced feature set $\mathcal{R} = \{\mathbf{R}_L^{vis}, \mathbf{R}_S^{vis}, \mathbf{R}_L^{ir}, \mathbf{R}_S^{ir}\}$, where each feature in $\mathcal{R}$ is obtained by Eq. (4).

Common Feature Space Learning (CFL) Module

Conventional feature-level VI-ReID methods usually learn a common feature space based on a contrastive-based loss, which directly minimizes the distances between VIS and IR features. However, such a manner cannot achieve a reasonable common feature space because of the large modality discrepancy. To address the above problem, it is desirable to learn the middle features from VIS and IR features, enabling us to obtain a reasonable common feature space.

A straightforward way to obtain a middle feature is to add or concatenate the VIS or IR features from different ranges. However, the above way cannot generate reliable middle features due to feature misalignment and loss of semantic information. Therefore, we propose a CFL module, which aligns the features from different modalities and ranges by graph attention (GAT) (Guo et al. 2021) and generates reliable middle features, as shown in Figure 1.

Specifically, we align each feature in $\mathcal{R}$ 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. First, we establish the similarity between $\mathbf{R}_L^{vis}$ and $\mathbf{R}_S^{ir}$ by using the inner product and the softmax function. This process can be formulated as

$$\mathbf{P} = \mathrm{Softmax}((\theta_q\mathbf{R}_L^{vis})(\theta_k\mathbf{R}_S^{ir})^{\mathrm{T}}), \tag{5}$$

where θ_q and θ_k are linear transformations; $\mathbf{P} \in \mathbb{R}^{N\times N}$ denotes the similarity matrix; and $\mathrm{Softmax}(\cdot)$ denotes the softmax function.

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

$$\begin{aligned}\bar{\mathbf{R}}_{ir,S}^{vis,L} &= \mathrm{GAT}(\mathbf{R}_L^{vis}, \mathbf{R}_S^{ir}) \\ &= \mathrm{ReLU}(\mathbf{P} - \lambda\mathrm{Mean}(\mathbf{P})\mathbf{1}\mathbf{1}^{\mathrm{T}})(\theta_v\mathbf{R}_S^{ir}),\end{aligned} \tag{6}$$

where $\mathrm{GAT}(\cdot)$ denotes the graph attention operation; θ_v 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 and the mean operation, respectively.

Based on the above, a middle feature $\mathbf{M}_L^{vis} \in \mathbb{R}^{N\times C}$ is obtained by aligning $\mathbf{R}_L^{vis}$ with $\mathbf{R}_S^{ir}$, $\mathbf{R}_L^{ir}$, and $\mathbf{R}_S^{vis}$, that is,

$$\begin{aligned}\mathbf{M}_L^{vis} = \mathrm{GAT}(\mathbf{R}_L^{vis}, \mathbf{R}_S^{ir}) + \mathrm{GAT}(\mathbf{R}_L^{vis}, \mathbf{R}_L^{ir}) + \\ \mathrm{GAT}(\mathbf{R}_L^{vis}, \mathbf{R}_S^{vis}) + \mathbf{R}_L^{vis}.\end{aligned} \tag{7}$$

Similar to Eq. (7), we can get the other reliable middle features. Hence, we obtain the middle feature set $\tilde{\mathcal{R}} =$

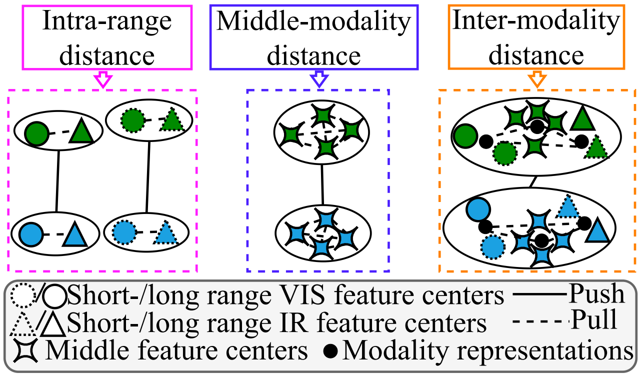

Figure 2: Illustration of the proposed MRIC loss. Different colors represent different identities.

$\{\mathbf{M}^{vis}_L, \mathbf{M}^{vis}_S, \mathbf{M}^{ir}_L, \mathbf{M}^{ir}_S\}$. To learn compact feature representations, following previous works (Ye et al. 2021b; Liu, Tan, and Zhou 2020), 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. In this way, we can get the 1D middle feature set $\tilde{\mathcal{R}}' = \{\mathbf{m}^{vis}_L, \mathbf{m}^{vis}_S, \mathbf{m}^{ir}_L, \mathbf{m}^{ir}_S\}$. Analogously, we apply the same pooling and concatenation operations to each feature in $\mathcal{R}$ and obtain 1D feature set $\mathcal{R}' = \{\mathbf{r}^{vis}_L, \mathbf{r}^{vis}_S, \mathbf{r}^{ir}_L, \mathbf{r}^{ir}_S\}$.

To mitigate the intra-class difference and inter-class discrepancy, we propose the MRIC loss to improve feature representations and reduce the distances between 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 the identity centers. The illustration of the MRIC loss is shown in Figure 2.

Technically, we first obtain identity centers, which are robust to pedestrian appearance changes, by the weighted average of the features of each person at one modality and a specific range. For instance, the center of the relation-enhanced features for the person with the identity i at the VIS modality and long-range can be obtained as

$$\mathbf{c}^{vis}_{L,i} = \sum_{j=1}^{K} \frac{\exp(\sum_{k=1}^{K} \mathbf{r}^{vis}_{L,i,j} {\mathbf{r}^{vis}_{L,i,k}}^{\mathrm{T}})}{\sum_{j=1}^{K} \exp(\sum_{k=1}^{K} \mathbf{r}^{vis}_{L,i,j} {\mathbf{r}^{vis}_{L,i,k}}^{\mathrm{T}})} \mathbf{r}^{vis}_{L,i,j}, \tag{8}$$

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

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

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

$$\mathcal{L}^{SL}_{MRIC} = \mathcal{L}^{\mathcal{C}^{vis}_S,\mathcal{C}^{ir}_S}_{MRIC} + \mathcal{L}^{\mathcal{C}^{vis}_L,\mathcal{C}^{ir}_L}_{MRIC}, \tag{9}$$

where

$$\begin{aligned} \mathcal{L}^{\mathcal{A},\mathcal{B}}_{MRIC} = & -\sum_{i=1}^{P} \log \frac{\exp(\mathbf{S}^{\mathcal{A},\mathcal{B}}_{i,i})}{\sum_{z=1}^{P}\exp(\mathbf{S}^{\mathcal{A},\mathcal{B}}_{i,z})} \\ & -\sum_{i=1}^{P} \log \frac{\exp(\mathbf{S}^{\mathcal{A},\mathcal{B}}_{i,i})}{\sum_{z=1}^{P}\exp(\mathbf{S}^{\mathcal{A},\mathcal{B}}_{z,i})} + \sum_{i=1}^{P} \mathcal{L}_1(\mathbf{a}_i - \mathbf{b}_i). \end{aligned} \tag{10}$$

Here, $\mathbf{S}^{\mathcal{A},\mathcal{B}} \in \mathbb{R}^{P\times P}$ denotes the cosine similarity matrix between $\mathcal{A}$ and $\mathcal{B}$ ($\mathbf{S}^{\mathcal{A},\mathcal{B}}_{i,j}$ denotes the cosine similarity between $\mathbf{a}_i$ (the i-th element of $\mathcal{A}$) and $\mathbf{b}_j$ (the j-th element of $\mathcal{B}$)); $\mathcal{L}_1(\cdot)$ represents the L_1 norm. By minimizing the first two terms, the similarities of the same person features are enhanced while those of the different person features are reduced. The last term denotes the L_1 distance between the same person features.

The middle-feature loss $\mathcal{L}^{MID}_{MRIC}$ is to reduce the distances between different middle features, that is,

$$\begin{aligned} \mathcal{L}^{MID}_{MRIC} = & \mathcal{L}^{\tilde{\mathcal{C}}^{vis}_S,\tilde{\mathcal{C}}^{vis}_L}_{MRIC} + \mathcal{L}^{\tilde{\mathcal{C}}^{vis}_S,\tilde{\mathcal{C}}^{ir}_S}_{MRIC} + \mathcal{L}^{\tilde{\mathcal{C}}^{vis}_S,\tilde{\mathcal{C}}^{ir}_L}_{MRIC} + \\ & \mathcal{L}^{\tilde{\mathcal{C}}^{vis}_L,\tilde{\mathcal{C}}^{ir}_S}_{MRIC} + \mathcal{L}^{\tilde{\mathcal{C}}^{vis}_L,\tilde{\mathcal{C}}^{ir}_L}_{MRIC} + \mathcal{L}^{\tilde{\mathcal{C}}^{ir}_S,\tilde{\mathcal{C}}^{ir}_L}_{MRIC}. \end{aligned} \tag{11}$$

The inter-modality loss $\mathcal{L}^{VIM}_{MRIC}$ is to reduce the intra-class distances and enlarge the inter-class distances between VIS, IR, and middle features, which is expressed as

$$\mathcal{L}^{VIM}_{MRIC} = \mathcal{L}^{\mathcal{C}^{vis},\mathcal{C}^{ir}}_{MRIC} + \mathcal{L}^{\mathcal{C}^{vis},\mathcal{C}^{mid}}_{MRIC} + \mathcal{L}^{\mathcal{C}^{ir},\mathcal{C}^{mid}}_{MRIC}, \tag{12}$$

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

Therefore, the MRIC loss is

$$\mathcal{L}_{MRIC} = \mathcal{L}^{SL}_{MRIC} + \mathcal{L}^{MID}_{MRIC} + \mathcal{L}^{VIM}_{MRIC}. \tag{13}$$

Joint Loss

The joint loss is defined as

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{TRI} + \mathcal{L}_{MRIC}, \tag{14}$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss and $\mathcal{L}_{TRI}$ denotes the triplet loss (Hermans, Beyer, and Leibe 2017).

The training of HOS-Net is given in *Supplement A*.

Experiments

Experimental Settings

Datasets. The SYSU-MM01 dataset (Wu et al. 2020) contains a total of 30,071 VIS images and 15,792 IR images from 491 different identities. The RegDB dataset (Nguyen et al. 2017) consists of 412 identities, where each identity has 10 VIS images and 10 IR images captured by two overlapping cameras. The LLCM dataset (Zhang and Wang

Methods	SYSU-MM01		RegDB		LLCM	
	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
D^2RL (Wang et al. 2019)	28.9 / 29.2	- / -	43.4 / 44.1	- / -	- / -	- / -
Hi-CMD (Choi et al. 2020)	34.9 / 35.9	- / -	70.9 / 66.0	- / -	- / -	- / -
JSIA-ReID (Wang et al. 2020)	38.1 / 36.9	43.8 / 52.9	48.1 / 48.9	48.5 / 49.3	- / -	- / -
X-Modality (Li et al. 2020)	49.9 / 50.7	- / -	62.2 / 60.2	- / -	- / -	- / -
DDAG (Ye et al. 2020)	54.8 / 53.0	61.0 / 68.0	69.3 / 63.5	68.1 / 61.8	48.0 / 52.3	40.3 / 48.4
LbA (Park et al. 2021)	55.4 / 54.1	58.5 / 66.3	74.2 / 67.6	67.5 / 72.4	50.8 / 55.6	43.8 / 53.1
G^2DA (Wan et al. 2023)	63.9 / 60.7	71.0 / 76.0	74.0 / 65.5	69.7 / 62.0	- / -	- / -
TSME (Liu et al. 2022b)	64.2 / 61.2	64.8 / 71.5	87.4 / 76.9	86.4 / 75.7	- / -	- / -
SPOT (Chen et al. 2022a)	65.3 / 62.3	69.4 / 74.6	80.4 / 72.5	79.4 / 72.3	- / -	- / -
PMT (Lu, Zou, and Zhang 2023)	67.5 / 65.0	71.7 / 76.5	84.8 / 76.6	84.2 / 75.1	- / -	- / -
CAJ (Ye et al. 2021a)	69.9 / 66.9	76.3 / 80.4	85.0 / 79.1	84.8 / 77.8	56.5 / 59.8	48.8 / 56.6
MMN (Zhang et al. 2021)	70.6 / 66.9	76.2 / 79.6	91.6 / 84.1	87.5 / 80.5	59.9 / 62.7	52.5 / 58.9
MAUM (Liu et al. 2022a)	71.7 / 68.8	77.0 / 81.9	87.9 / 85.1	87.0 / 84.3	- / -	- / -
DEEN (Zhang and Wang 2023)	74.7 / 71.8	80.3 / 83.3	91.1 / 85.1	89.5 / 83.4	62.5 / 65.8	54.9 / 62.9
HOS-Net (Ours)	**75.6 / 74.2**	**84.2 / 86.7**	**94.7 / 90.4**	**93.3 / 89.2**	**64.9 / 67.9**	**56.4 / 63.2**

Table 1: Comparisons with state-of-the-art methods on the SYSU-MM01, RegDB and LLCM datasets. The bold font and the underline denote the best and second-best performance, respectively.

2023) is captured in low-light environments. The training set contains 713 identities (with 16,946 VIS images and 13,975 IR images) while the test set contains 351 identities (with 8,680 VIS images and 7,166 IR images).

Implementation Details. All the images are resized to 256×128 with horizontal flip, random erasing, and channel augmentation for data augmentation (Ye et al. 2021a) during the training phase. For each mini-batch, we randomly choose 8 identities, where 4 VIS images and 4 IR images of each identity are selected. We adopt AGW (Ye et al. 2021b) as our backbone. We use the warm-up strategy to update the learning rate from 0.01 to 0.1 at the first 10 epochs. At the 20 and 50 epochs, the learning rates are set to 0.01 and 0.001, respectively. We use SGD as the optimizer and the momentum parameter is set to 0.9. The total number of training epochs is set to 120. Our proposed HOS-Net is implemented with the PyTorch on an NVIDIA RTX3090 GPU. The number of hyperedges M is set to 256. λ in Eq. (6) is set to 1.3.

Cumulative Matching characteristics (CMC) and mean Average Precision (mAP) are used as our evaluation metrics. CMC measures the matching probability of the ground-truth person occurring in the top-k retrieved results (Rank-k accuracy). Besides, we randomly divide the RegDB dataset for training and testing. The above process is repeated ten times and we report the average performance. We also randomly split the gallery set of the SYSU-MM01 and LLCM datasets ten times to report the average performance.

Comparison with State-of-the-Art Methods

The comparison results are given in Table 1. More results are shown in *Supplement B*.

SYSU-MM01. As shown in Table 1, our proposed HOS-Net obtains the best or comparable performance among all the competing methods. Specifically, HOS-Net gives about 13.0% and 15.2% improvements in terms of mAP over some image-level methods (such as JSIA-ReID and TSME) for both all and indoor search modes, respectively. Compared with the CNN-based method (DDAG) and Transformer-based method (PMT), HOS-Net surpasses them by at least 8.1% in Rank-1 and 9.2% in mAP for the all search mode. Moreover, for the indoor search mode, HOS-Net outperforms the second-best method DEEN by 3.9% in Rank-1 and 3.4% in mAP. DEEN ignores the importance of high-order structure information, leading to inferior performance.

RegDB. From Table 1, we can also observe that our proposed HOS-Net achieves the best performance for two search modes. For two search modes, our HOS-Net outperforms MMN by 3.1%/6.3% and 5.8%/8.7% in Rank-1/mAP, respectively. Moreover, compared with G^2DA and SPOT, which rely on high-quality person structure labels to obtain modality-shared features, HOS-Net improves the Rank-1 and mAP by at least 13.9% and 16.9%, respectively, for the IR to VIS search mode. This further indicates the superiority of our high-order structure-based network for VI-ReID.

LLCM. We also report the comparison results on the LLCM dataset in Table 1. For the IR to VIS search mode, our HOS-Net outperforms MMN by 3.9% and 4.3% in terms of Rank-1 and mAP, respectively. Moreover, HOS-Net performs significantly better than the second-best DEEN for the VIS to IR search mode, achieving the best results with 64.9%/67.9% in Rank-1/mAP. Therefore, HOS-Net can learn a discriminative and reasonable common feature space to reduce the modality discrepancy.

Ablation Studies

Effectiveness of Key Components. We conduct ablation studies to validate the effectiveness of each key component of the proposed HOS-Net (including SLE, HSL, CFL, and the MRIC loss). The results are shown in Table 2, where Method 1 represents the baseline AGW method.

SLE: By introducing SLE, Method 2 achieves about 2.5% and 5.7% higher mAP than Method 1 on the SYSU-MM01 and RegDB datasets, respectively. This shows the effectiveness of our SLE, which explores different ranges of person

#	Methods	SYSU-MM01	RegDB
		R-1 / mAP	R-1 / mAP
1	Baseline	69.9 / 66.9	85.0 / 79.1
2	Baseline+SLE	71.7 / 69.4	89.6 / 84.8
3	+HSL	73.3 / 72.4	92.0 / 87.1
4	+CFL	72.1 / 70.2	91.6 / 86.5
5	+HSL+CFL	74.0 / 72.9	92.7 / 87.8
6	+CFL +$\mathcal{L}_{MRIC}$	74.5 / 72.7	93.2 / 88.4
7	+HSL+CFL +$\mathcal{L}_{MRIC}$	**75.6 / 74.2**	**94.7 / 90.4**

Table 2: The influence of key components of HOS-Net on the performance on the SYSU-MM01 and RegDB datasets.

Settings		SYSU-MM01	RegDB
Hypergraph	Whitening	R-1 / mAP	R-1 / mAP
-	-	71.7 / 69.4	89.6 / 84.8
✓	-	72.5 / 70.3	91.1 / 86.3
✓	✓	**73.3 / 72.4**	**92.0 / 87.1**

Table 3: The influence of the hypergraph and the whitening operation on the SYSU-MM01 and RegDB datasets.

Settings			SYSU-MM01
	Modalities	Ranges	R-1 / mAP
Addition	✓	-	72.0 / 71.1
	-	✓	71.8 / 70.9
	✓	✓	71.4 / 70.4
Concatenation	✓	-	71.9 / 70.8
	-	✓	72.3 / 71.1
	✓	✓	72.5 / 71.2
GAT	✓	-	73.4 / 72.3
	-	✓	73.6 / 72.4
	✓	✓	**74.0 / 72.9**

Table 4: The influence of generating middle features from different modality and range features on SYSU-MM01.

Settings	SYSU-MM01	RegDB
	R-1 / mAP	R-1 / mAP
-	74.0 / 72.9	92.7 / 87.8
+$\mathcal{L}_{MRIC}^{SL}$	74.3 / 73.3	93.4 / 88.6
+$\mathcal{L}_{MRIC}^{MID}$	74.4 / 73.2	93.3 / 88.3
+$\mathcal{L}_{MRIC}^{SL}$ +$\mathcal{L}_{MRIC}^{MID}$	75.0 / 73.8	93.8 / 89.2
+$\mathcal{L}_{MRIC}^{SL}$ +$\mathcal{L}_{MRIC}^{MID}$+$\mathcal{L}_{MRIC}^{VIM}$	**75.6 / 74.2**	**94.7 / 90.4**

Table 5: The influence of each term in the MRIC loss.

features by taking advantage of both CNN and Transformer. **HSL:** By incorporating HSL into Method 2, Method 3 achieves 1.6%/3.0% and 2.4%/2.3% improvements in Rank-1/mAP on two datasets, respectively. This validates the importance of HSL, which adopts the whitened hypergraph network to model the high-order relationship across different local features of each person image and avoid model collapse. **CFL:** Method 5 introduces CFL to Method 3 and it obtains higher accuracy (0.7%/0.7% improvements in Rank-1/mAP on the RegDB dataset) than Method 3. This demonstrates that learning reliable middle features can effectively reduce the modality discrepancy. **The MRIC loss:** Compared with Method 5, Method 7 achieves 1.6%/1.3% and 2.0%/2.6% improvements in Rank-1/mAP on two datasets, respectively. The MRIC loss can improve feature representations and reduce discrepancies between the VIS and IR modalities, achieving a reasonable common feature space.

Effectiveness of the Hypergraph and the Whitening Operation. HSL is based on a whitened hypergraph network to discover the high-order relationship of person features and avoid model collapse. As shown in Table 3, by modeling the high-order structure with the hypergraph, the model brings about 0.8%/0.9% gains in Rank-1/mAP on SYSU-MM01. Note that the original hypergraph network allows unrestricted connections among nodes to capture high-order structure information, suffering from model collapse during hypergraph learning. By adding the whitening operation into the hypergraph learning, the performance is improved by 0.8%/2.1% and 0.9%/0.8% in Rank-1/mAP on two datasets, respectively. Hence, the whitening operation is beneficial to alleviate mode collapse and improve the final performance.

Influence of Different Middle Features. The CFL module leverages graph attention to align features from different modalities and ranges to generate reliable middle features. In this subsection, we evaluate the influence of different middle features on the performance. The results are given in Table 4. Compared with the methods that generate middle features by adding or concatenating the VIS or IR features, our method with GAT improves mAP on the SYSU-MM01 datasets, respectively. This clearly indicates the effectiveness of the middle features generated by GAT.

Influence of Each Term in the MRIC Loss. The MRIC loss is proposed to improve feature representations and reduce the distances between the VIS, IR, and middle features. As shown in Table 5, when all the terms in the MRIC loss are used to jointly train the network, Rank-1/mAP is improved by 1.6%/1.3% and 2.0%/2.6% in comparison with HOS-Net trained without the MRIC loss on two datasets, respectively. This indicates that HOS-Net trained with the MRIC loss can achieve a reasonable common feature space.

More ablation studies and visualization results can refer to *Supplement C and D*.

Conclusion

In this paper, we propose a novel HOS-Net consisting of the backbone, SLE, HSL, and CFL modules for VI-ReID. The SLE module is first designed to learn short-range and long-range features by taking advantage of both CNN and Transformer. Then, the HSL module exploits diverse high-order structure information of features without suffering from model collapse based on a whitened hypergraph. Finally, the CFL module generates reliable middle features and obtains a reasonable common feature space. Extensive experiments on three VI-ReID benchmarks verify the effectiveness of HOS-Net in comparison with several state-of-the-art methods. Currently, the training complexity of our method is still high (see *Supplement C* for more details). We plan to explore new ways to reduce the training complexity.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62372388 and 62071404, by the Open Research Projects of Zhejiang Lab under Grant 2021KG0AB02, and by the Natural Science Foundation of Fujian Province under Grant 2021J011185.

References

Chen, C.; Ye, M.; Qi, M.; Wu, J.; Jiang, J.; and Lin, C.-W. 2022a. Structure-aware positional Transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 31: 2352–2364.

Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; and Liu, Z. 2022b. Mobile-former: Bridging MobileNet and Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5270–5279.

Choi, S.; Lee, S.; Kim, Y.; Kim, T.; and Kim, C. H.-C. 2020. Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13–19.

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-modality person re-identification with generative adversarial training. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1–7.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3558–3565.

Guo, D.; Shao, Y.; Cui, Y.; Wang, Z.; Zhang, L.; and Shen, C. 2021. Graph attention tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9543–9552.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Higham, D. J.; and de Kergorlay, H.-L. 2022. Mean field analysis of hypergraph contagion models. *SIAM Journal on Applied Mathematics*, 82(6): 1987–2007.

Huang, N.; Liu, J.; Luo, Y.; Zhang, Q.; and Han, J. 2023. Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification. *Pattern Recognition*, 135: 109145.

Huang, N.; Liu, K.; Liu, Y.; Zhang, Q.; and Han, J. 2022. Cross-modality person re-identification via multi-task learning. *Pattern Recognition*, 128: 108653.

Jin, X.; He, T.; Zheng, K.; Yin, Z.; Shen, X.; Huang, Z.; Feng, R.; Huang, J.; Chen, Z.; and Hua, X. 2022. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14278–14287.

Li, D.; Wei, X.; Hong, X.; and Gong, Y. 2020. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4610–4617.

Li, Y.; Guo, Z.; Zhang, H.; Li, M.; and Ji, G. 2021. Decoupled pose and similarity based graph neural network for video person re-identification. *IEEE Signal Processing Letters*, 29: 264–268.

Liu, H.; Tan, X.; and Zhou, X. 2020. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Transactions on Multimedia*, 23: 4414–4425.

Liu, J.; Sun, Y.; Zhu, F.; Pei, H.; Yang, Y.; and Li, W. 2022a. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19366–19375.

Liu, J.; Wang, J.; Huang, N.; Zhang, Q.; and Han, J. 2022b. Revisiting modality-specific feature compensation for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10): 7226–7240.

Lu, H.; Zou, X.; and Zhang, P. 2023. Learning progressive modality-shared Transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1835–1843.

Lu, J.; Wan, H.; Li, P.; Zhao, X.; Ma, N.; and Gao, Y. 2023. Exploring high-order spatio-temporal correlations from skeleton for person re-identification. *IEEE Transactions on Image Processing*, 32: 949–963.

Nguyen, D. T.; Hong, H. G.; Kim, K. W.; and Park, K. R. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3): 605.

Park, H.; Lee, S.; Lee, J.; and Ham, B. 2021. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12046–12055.

Wan, L.; Sun, Z.; Jing, Q.; Chen, Y.; Lu, L.; and Li, Z. 2023. D^{2}DA: Geometry-guided dual-alignment learning for RGB-infrared person re-identification. *Pattern Recognition*, 135: 109150.

Wang, G.; Zhang, T.; Yang, Y.; Cheng, J.; Chang, J.; Liang, X.; and Hou, Z. 2020. Cross-modality paired-images generation for RGB-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12144–12151.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.; and Satoh, S. 2019. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 618–626.

Wei, Z.; Yang, X.; Wang, N.; and Gao, X. 2022. Rbdf: Reciprocal bidirectional framework for visible infrared person reidentification. *IEEE Transactions on Cybernetics*, 52(10): 10988–10998.

Wu, A.; Zheng, W.; Gong, S.; and Lai, J. 2020. RGB-IR person re-identification by cross-modality similarity preservation. *International Journal of Computer Vision*, 128: 1765–1785.

Yan, C.; Pang, G.; Jiao, J.; Bai, X.; Feng, X.; and Shen, C. 2021. Occluded person re-identification with single-scale global representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11875–11884.

Yang, B.; Chen, J.; and Ye, M. 2023. Top-K Visual Tokens Transformer: Selecting Tokens for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.

Ye, M.; Ruan, W.; Du, B.; and Shou, M. Z. 2021a. Channel Augmented Joint Learning for Visible-Infrared Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13567–13576.

Ye, M.; Shen, J.; J. Crandall, D.; Shao, L.; and Luo, J. 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Proceedings of the European Conference on Computer Vision*, 229–247.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021b. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 2872–2893.

Zhang, D.; Zhang, Z.; Ju, Y.; Wang, C.; Xie, Y.; and Qu, Y. 2022. Dual mutual learning for cross-modality person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8): 5361–5373.

Zhang, H.; Hu, W.; and Wang, X. 2022. Parc-Net: Position aware circular convolution with merits from convnets and Transformer. In *Proceedings of the European Conference on Computer Vision*, 613–630.

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

Zhang, Y.; Yan, Y.; Lu, Y.; and Wang, H. 2021. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the ACM International Conference on Multimedia*, 788–796.