Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hierarchical Identity-Constrained Representation Learning with Statistical Calibration for Visible–Infrared Person Re-Identification

Rongyu Jia
Guizhou Normal University
Haiyang Chang
Guizhou Normal University
Daoxun Xia
dxxia@gznu.edu.cn

Guizhou Normal University

Research Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10559216/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hierarchical Identity-Constrained Representation Learning with Statistical Calibration for Visible–Infrared Person Re-Identification

Rongyu Jia[1], Haiyang Chang[1], Daoxun Xia[1,2*]

[1]Guizhou Key Laboratory of Advanced Computing
Guizhou Normal University, Guiyang, 550025, China.
[2*]Leading Talent Workstation for Technological Innovation
in Artificial Intelligence of Guizhou Province
Guizhou Normal University, Guiyang, 550025, China.

*Corresponding author(s). E-mail(s): dxxia@gznu.edu.cn;
Contributing authors: 1173753172@qq.com; 2041574449@qq.com;

Abstract

Visible–infrared person re-identification (VI-ReID) seeks to match identities across heterogeneous modalities, yet modality discrepancy perturbs intermediate feature statistics and shifts high-level identity semantics. Existing methods often emphasize one of these effects, while their coordinated treatment across the feature hierarchy remains underexplored. We propose a hierarchical identity-constrained representation learning framework with statistical calibration, termed HICR. HICR integrates Statistical Modulation Feature Aggregation (SMFA) and Identity Conservation Learning (ICL). SMFA recalibrates modality-sensitive intermediate responses through variance-guided modulation and lightweight local processing, while ICL regularizes identity semantics across high-level descriptor spaces using adaptive prototypes, prototype residual consistency, and bidirectional cross-group prototype classification. Joint optimization connects intermediate statistical calibration with high-level semantic regularization. Experiments on SYSU-MM01, RegDB, and LLCM demonstrate competitive performance. Controlled ablations support the effectiveness and complementarity of SMFA and ICL, while efficiency and representation analyses further characterize the design. Multi-seed evaluations indicate reproducible gains for the individual components.

Keywords: visible–infrared person re-identification; cross-modal representation learning; statistical calibration; prototype learning; identity-semantic consistency

1 Introduction

Visible-infrared person re-identification (VI-ReID) aims to establish reliable pedestrian identity correspondence across heterogeneous imaging modalities and plays an important role in around-the-clock intelligent video surveillance [1, 2]. Compared with conventional single-modality person re-identification, VI-ReID must additionally address the substantial modality discrepancy caused by different imaging mechanisms. Visible images mainly capture appearance information formed by reflected light, whereas infrared images emphasize thermal responses. Consequently, the same pedestrian may exhibit substantially different visual characteristics across the two modalities, making cross-modal identity matching particularly challenging.

The modality discrepancy is not confined to the final shared embedding space but may progressively propagate through feature extraction. At intermediate layers, it can appear as shifts in channel statistics, response dispersion, and unstable local activation patterns. In high-level representation spaces, it may further cause identity-semantic center shifts across modalities or descriptor groups. The former weakens the representation basis required by subsequent identity supervision, whereas the latter directly limits consistent cross-modal identity aggregation. VI-ReID therefore requires not only discriminative features but also coordinated treatment of intermediate statistical perturbations and high-level identity-semantic shifts.

Existing studies address modality discrepancy through modality compensation [3, 4], structural modeling [5, 6], semantic consistency [7, 8], and center- or class-level constraints [9, 10]. Although some methods include architectural and objective components at both levels, existing approaches are often dominated by either representation-level enhancement or objective-level semantic regularization. The coordinated treatment of intermediate statistical calibration and high-level identity-semantic regularization across the feature hierarchy remains underexplored. Representation-level methods improve cross-modal feature quality, but stronger features alone do not guarantee consistent semantic centers for the same identity across multiple high-level descriptor spaces. Conversely, high-level semantic objectives improve identity-space organization but generally rely on sufficiently stable upstream representations. When intermediate features still contain modality-induced statistical perturbations, identity-level supervision may be imposed on an unreliable representation basis.

To address this gap, we propose a Hierarchical Identity-Constrained Representation (HICR) learning framework with statistical calibration for VI-ReID. As illustrated in Fig. 1, HICR addresses modality discrepancy at two representation levels through Statistical Modulation Feature Aggregation (SMFA) and Identity Conservation Learning (ICL). SMFA employs channel-wise variance as a statistical prior to modulate modality-sensitive intermediate responses and combines this operation with lightweight local processing and residual aggregation. On this recalibrated representation basis, ICL independently constructs consistency-aware identity prototypes within multiple high-level descriptor groups and regularizes valid group pairs through prototype residual consistency and bidirectional cross-group prototype classification. Together, SMFA and ICL establish a hierarchical learning process from intermediate statistical calibration to high-level identity-semantic regularization.

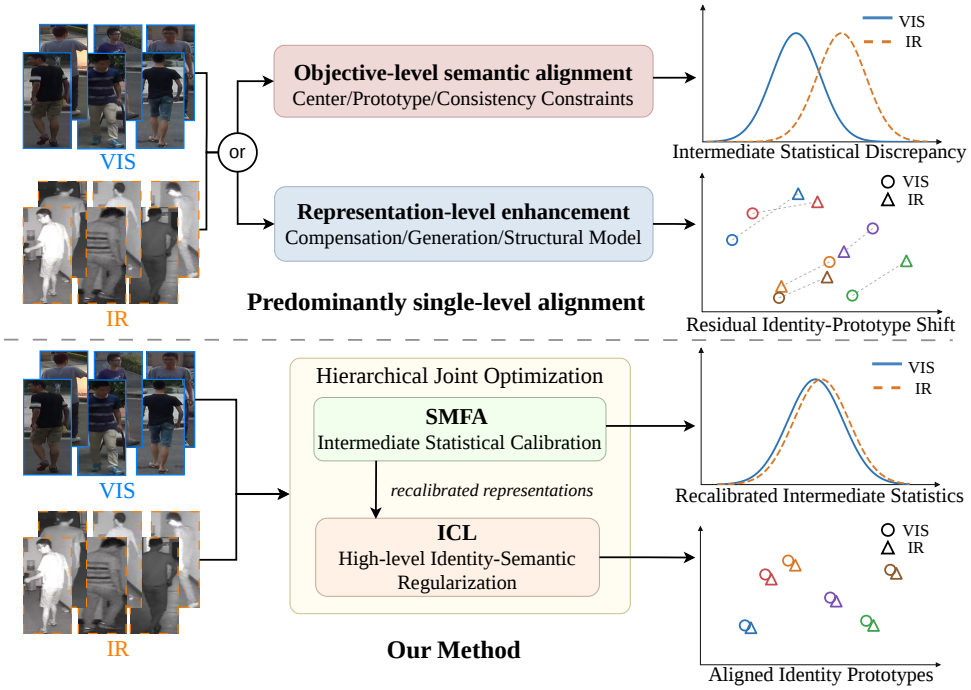


Fig. 1: Conceptual comparison between predominantly single-level alignment and the proposed hierarchical learning strategy. Although some existing methods contain components at both the representation and objective levels, they are often dominated by either representation enhancement or semantic regularization, potentially leaving intermediate statistical perturbations or residual identity-prototype shifts. HICR couples SMFA-based intermediate statistical modulation with ICL-based high-level identity-semantic regularization to jointly address the two forms of cross-modal discrepancy.

The main contributions of this work are summarized as follows:

- We propose HICR, which hierarchically couples intermediate statistical calibration with high-level identity-semantic constraints to address two complementary manifestations of modality discrepancy.
- We develop SMFA, which uses channel-wise variance as a statistical prior for global feature-response modulation and combines it with lightweight local processing and residual aggregation to recalibrate modality-sensitive intermediate representations.
- We introduce ICL, which enhances same-identity semantic consistency and inter-identity discrimination across multiple high-level descriptor spaces through consistency-aware adaptive prototypes, prototype residual consistency, and bidirectional cross-group prototype classification.
- Experiments on SYSU-MM01, RegDB, and LLCM evaluate HICR across multiple benchmarks and retrieval directions. Controlled component and internal ablations

examine the contributions and complementarity of SMFA and ICL. Parameter-sensitivity, insertion-position, efficiency, and representation analyses further characterize the proposed design, while multi-seed evaluations indicate reproducible gains for the individual components.

2 Related Work

This section reviews related VI-ReID research from three complementary perspectives: general VI-ReID, representation-level cross-modal learning, and objective-level semantic regularization. These categories are organized according to the primary modeling target of each line of work rather than treated as mutually exclusive. For each category, we summarize its main objective, representative developments, remaining capability boundary, and relation to HICR.

2.1 Visible-Infrared Person Re-Identification

Visible-infrared person re-identification aims to match pedestrian identities across visible and infrared modalities. Owing to their heterogeneous imaging mechanisms, visible images mainly capture reflected appearance information, whereas infrared images emphasize thermal responses. Cross-modal representations are therefore affected by appearance discrepancies, statistical shifts during feature extraction, and identity-semantic inconsistency in high-level embedding spaces [11, 12].

Existing studies can be broadly organized according to two primary modeling targets. Representation-level methods improve cross-modal features through modality compensation, image generation, middle-modality construction, and structural enhancement [3, 5, 6, 13–16]. Objective-level methods organize the shared embedding space using center constraints, semantic consistency, class-level supervision, and prototype learning [7–10, 17, 18].

These two directions are complementary rather than mutually exclusive. Representation-level methods mainly focus on cross-modal feature formation and enhancement, whereas objective-level methods primarily optimize identity distributions. However, their interaction across the feature hierarchy is often not explicitly modeled. HICR is positioned at the intersection of these directions by linking SMFA-based intermediate statistical calibration with ICL-based high-level identity-semantic regularization.

2.2 Representation-Level Cross-Modal Learning

Representation-level methods reduce visible-infrared discrepancy during input transformation or feature extraction. Early studies adopted modality compensation [3, 4, 19], cross-modal correspondence learning [20], and intermediate-modality or generation-based alignment [13, 14], while auxiliary structural cues were also explored to facilitate cross-modal correspondence [21].

Recent methods further employ structure-aware and geometry-guided modeling [5, 6, 15], part- and local-detail learning [16, 22, 23], spectrum-aware augmentation and attention-guided correlation learning [24, 25], and disentangled, frequency-aware,

or Transformer-based representations [26–29]. These strategies substantially improve feature discriminability and cross-modal robustness.

However, their primary objective is typically to generate, enhance, or align stronger cross-modal features rather than explicitly characterize how modality-induced perturbations in channel means, variances, and local responses evolve across intermediate layers. Residual statistical instability may therefore remain in the feature hierarchy and weaken the representation basis for subsequent identity supervision. Moreover, stronger representations do not necessarily ensure that the same identity forms consistent semantic centers across modalities or multiple high-level descriptor spaces. This capability boundary motivates SMFA as an intermediate response-calibration module, while identity-level semantic consistency is addressed at a higher representation level.

2.3 Objective-Level Semantic Regularization and Prototype Learning

Objective-level methods directly organize the shared embedding space through training objectives. Identity classification and triplet loss are commonly used to improve intra-identity compactness and inter-identity separation. Center-based constraints and heterogeneous center modeling further reduce cross-modal identity shifts [9, 10], while auxiliary modality learning and memory mechanisms provide additional supervision [30–32].

Recent studies also explore semantic consistency [7, 8, 13, 33], class-aware supervision [18], dual-space alignment [17], channel-level augmentation [34], and privileged or multi-stage auxiliary learning [35, 36]. Compared with purely instance-level constraints, these methods provide more stable identity-level guidance.

Despite these advances, several aspects remain insufficiently characterized. Many objectives mainly model instance-to-center or sample-to-class relations, whereas identity correspondence among multiple high-level descriptor groups is less explicitly considered. Uniformly averaged prototypes may be sensitive to noisy, occluded, or less representative samples, and reducing same-identity center distances alone does not necessarily preserve inter-identity boundaries. In addition, high-level semantic objectives are generally applied on top of learned representations; when upstream features remain statistically unstable, the resulting identity-level regularization may be less reliable.

In summary, representation-level methods mainly improve cross-modal feature formation but do not necessarily guarantee high-level identity-semantic consistency, whereas objective-level methods organize identity distributions and benefit from a stable intermediate representation basis. HICR connects these complementary directions by coupling SMFA-based intermediate statistical calibration with ICL-based identity-prototype regularization across multiple high-level descriptor spaces.

3 Proposed Method

This section presents HICR, the proposed hierarchical identity-constrained representation learning framework with statistical calibration. HICR integrates Statistical Modulation Feature Aggregation (SMFA) within the shared backbone with Identity

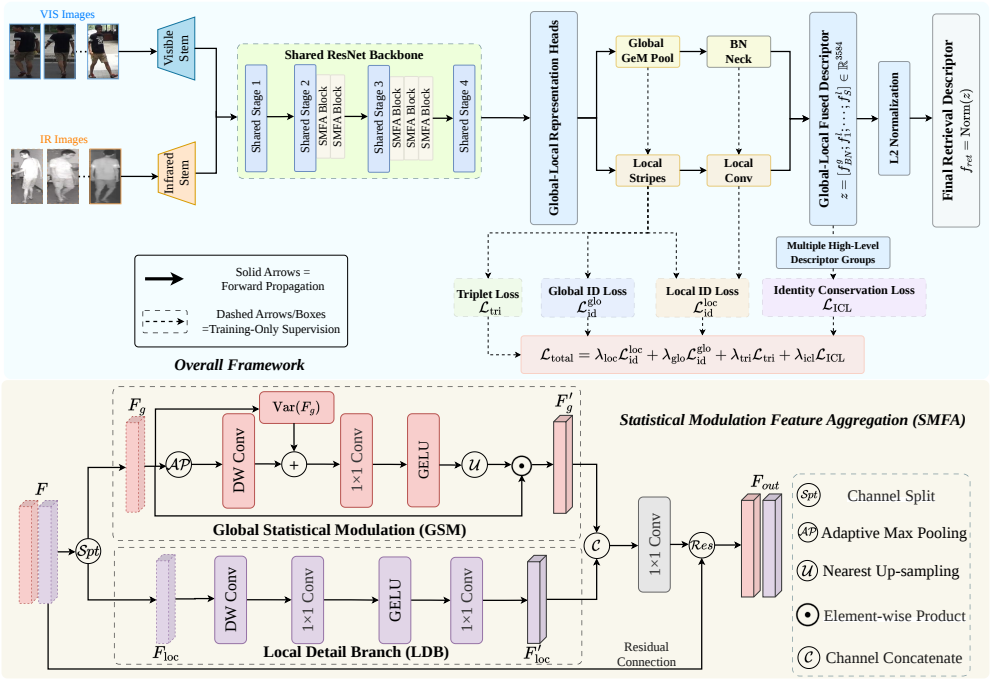


Fig. 2: Overview of HICR and SMFA. Visible and infrared images are processed by modality-specific stems and a parameter-shared backbone equipped with SMFA blocks, followed by global-local representation heads. During training, the resulting high-level representations are organized into multiple fused descriptor groups for ICL, whereas only the L_2-normalized global-local fused descriptor is used for cross-modal retrieval during inference. Dashed arrows and boxes denote training-only supervision. The lower panel illustrates GSM, LDB, and residual aggregation in SMFA.

Conservation Learning (ICL) over multiple high-level descriptor groups. SMFA recalibrates modality-sensitive intermediate feature responses, whereas ICL regularizes identity semantics across high-level descriptor spaces, forming a hierarchical learning process from intermediate statistical calibration to high-level identity-semantic regularization.

3.1 Framework Overview

As illustrated in Fig. 2, HICR consists of modality-specific shallow stems, a parameter-shared backbone equipped with SMFA blocks, and global and local representation heads. Let $(\mathbf{x}_i^m, y_i)$ denote a training sample, where $\mathbf{x}_i^m$ is the i-th person image from modality $m \in \{v, r\}$, v and r denote the visible and infrared modalities, respectively, and y_i is the identity label. The two modalities are first processed by their respective shallow stems, yielding $\mathbf{F}_{i,0}^m = \mathcal{S}_m(\mathbf{x}_i^m)$, and are then fed into the parameter-shared backbone.

To alleviate modality-induced perturbations in channel statistics and local responses, SMFA blocks are inserted into selected shared residual units. The feature extraction process at the l-th residual unit is formulated as

$$\mathbf{F}_{i,l}^{m} = \begin{cases} \mathcal{M}_l\Big(\mathcal{B}_l(\mathbf{F}_{i,l-1}^{m})\Big), & l \in \mathcal{I}, \\ \mathcal{B}_l(\mathbf{F}_{i,l-1}^{m}), & l \notin \mathcal{I}, \end{cases} \qquad l = 1, \ldots, L. \tag{1}$$

Here, $\mathcal{B}_l$ denotes the l-th shared residual transformation, $\mathcal{M}_l$ denotes the corresponding SMFA mapping, and $\mathcal{I}$ is the set of SMFA insertion positions.

The final backbone feature $\mathbf{F}_{i,L}^{m}$ is further encoded by global and local representation heads. The global branch employs generalized mean (GeM) pooling and a BN neck to capture holistic identity semantics, whereas the local branch uses horizontal stripe partitioning and local transformations to extract complementary fine-grained cues:

$$\mathbf{z}_{i,\mathrm{glo}}^{m} = \mathcal{H}_{\mathrm{glo}}\left(\mathbf{F}_{i,L}^{m}\right), \qquad \mathbf{z}_{i,\mathrm{loc}}^{m} = \mathcal{H}_{\mathrm{loc}}\left(\mathbf{F}_{i,L}^{m}\right). \tag{2}$$

During training, the high-level outputs are organized into multiple fused descriptor groups,

$$\mathcal{Z} = \left\{\mathbf{Z}^{(q)}\right\}_{q=1}^{Q}, \tag{3}$$

where $\mathbf{Z}^{(q)}$ denotes the q-th high-level descriptor group. ICL selects valid group pairs from $\mathcal{Z}$ and imposes prototype-level identity constraints across the paired descriptor spaces, as detailed in Section 3.3. During inference, the global and local representations are concatenated and L_2-normalized to obtain the final retrieval descriptor:

$$\mathbf{z}_{i}^{m} = \mathrm{Cat}\left(\mathbf{z}_{i,\mathrm{glo}}^{m}, \mathbf{z}_{i,\mathrm{loc}}^{m}\right), \qquad \mathbf{f}_{i,\mathrm{ret}}^{m} = \frac{\mathbf{z}_{i}^{m}}{\|\mathbf{z}_{i}^{m}\|_2}. \tag{4}$$

The network is jointly optimized using global and local identity classification losses, triplet loss, and ICL. SMFA recalibrates intermediate feature responses within the shared backbone, whereas ICL regularizes identity semantics across the training-time descriptor groups. Only the normalized descriptor $\mathbf{f}_{i,\mathrm{ret}}^{m}$ is retained for cross-modal similarity matching during inference.

3.2 Statistical Modulation Feature Aggregation

The heterogeneous imaging mechanisms of visible and infrared cameras may propagate modality discrepancy into intermediate layers, where it appears as shifts in channel responses and response dispersion. As illustrated in the lower panel of Fig. 2, SMFA uses statistical priors to adaptively recalibrate intermediate features. For notational simplicity, the sample index i is omitted. Let $\mathbf{F}_{l}^{m} \in \mathbb{R}^{C_l \times H_l \times W_l}$ denote the feature map at layer l from modality $m \in \{v, r\}$. The spatial mean and variance of its c-th channel are defined as

$$\mu_{l}^{m}(c) = \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \mathbf{F}_{l}^{m}(c, h, w). \tag{5}$$

$$\sigma_l^{m,2}(c) = \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left[\mathbf{F}_l^m(c,h,w) - \mu_l^m(c)\right]^2. \tag{6}$$

The channel statistics are collected as $\boldsymbol{\mu}_l^m = [\mu_l^m(1), \ldots, \mu_l^m(C_l)]^\top$ and $\boldsymbol{\sigma}_l^{m,2} = [\sigma_l^{m,2}(1), \ldots, \sigma_l^{m,2}(C_l)]^\top$. The cross-modal statistical discrepancy at layer l is characterized as

$$\Delta_{\text{stat}}^l = \|\boldsymbol{\mu}_l^v - \boldsymbol{\mu}_l^r\|_2 + \left\|\boldsymbol{\sigma}_l^{v,2} - \boldsymbol{\sigma}_l^{r,2}\right\|_2. \tag{7}$$

This quantity provides an analytical description of modality-related perturbations from first- and second-order channel statistics and is not directly optimized as an additional loss. SMFA instead employs channel-wise variance as a statistical prior for feature-response modulation.

At an SMFA insertion position defined in Eq. (1), let $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ denote the input feature. SMFA evenly splits $\mathbf{F}$ along the channel dimension into a global statistical branch and a local detail branch:

$$\mathbf{F} = \text{Cat}\left(\mathbf{F}_g, \mathbf{F}_{\text{loc}}\right), \qquad \mathbf{F}_g, \mathbf{F}_{\text{loc}} \in \mathbb{R}^{C_g\times H\times W}, \qquad C_g = \frac{C}{2}. \tag{8}$$

The global branch estimates the statistical prior and performs feature-response modulation, whereas the local branch provides a complementary local processing pathway.

In the global statistical modulation branch, a compact spatial response and a channel-wise variance prior are estimated in parallel:

$$\mathbf{Q}_g = \text{DWConv}\left[\text{Pool}_{8\times 4}\left(\mathbf{F}_g\right)\right], \qquad \mathbf{v}_g = \text{Var}_{H,W}\left(\mathbf{F}_g\right), \tag{9}$$

where $\mathbf{Q}_g \in \mathbb{R}^{C_g\times 8\times 4}$ and $\mathbf{v}_g \in \mathbb{R}^{C_g\times 1\times 1}$. The compact response $\mathbf{Q}_g$ provides coarse spatial context, whereas $\mathbf{v}_g$ describes channel-response dispersion. After broadcasting the variance prior, the two terms are combined to generate a modulation map:

$$\mathbf{M}_g = \text{Up}_{\text{NN}}\left\{\phi\left[\text{Conv}_{1\times 1}\left(\mathbf{Q}_g + \text{Broadcast}(\mathbf{v}_g)\right)\right]\right\}. \tag{10}$$

The global feature is then modulated through element-wise multiplication:

$$\widetilde{\mathbf{F}}_g = \mathbf{F}_g \odot \mathbf{M}_g. \tag{11}$$

Here, $\phi(\cdot)$ denotes the GELU activation, $\text{Up}_{\text{NN}}(\cdot)$ denotes nearest-neighbor upsampling to the original spatial resolution $H \times W$, and $\odot$ denotes element-wise multiplication. This branch uses both spatial context and response dispersion to adaptively modulate intermediate features.

The local detail branch applies a depthwise convolution followed by two pointwise transformations to the complementary channel subset:

$$\widetilde{\mathbf{F}}_{\text{loc}} = \text{Conv}_{1\times 1}^{(2)}\left\{\phi\left[\text{Conv}_{1\times 1}^{(1)}\left(\text{DWConv}\left(\mathbf{F}_{\text{loc}}\right)\right)\right]\right\}. \tag{12}$$

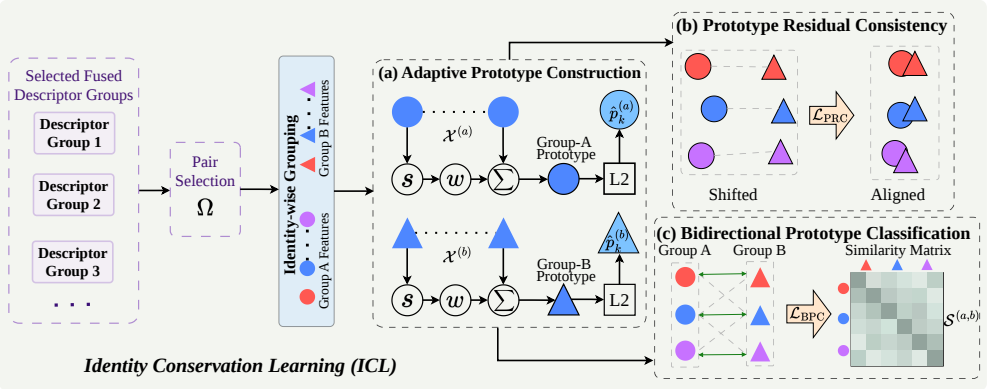


Fig. 3: Illustration of ICL for an arbitrary valid pair selected from multiple high-level descriptor groups. Samples are first organized identity-wise and aggregated into consistency-weighted identity prototypes. Prototype residual consistency reduces same-identity discrepancy across the paired spaces, while bidirectional prototype classification preserves inter-identity discrimination through cross-group matching.

Unlike GSM, this branch does not explicitly use a statistical prior; instead, it provides lightweight local processing based on neighborhood information.

The two branch outputs are concatenated and projected to a residual calibration term:

$$\mathbf{F}_c = \mathrm{Cat}\left(\widetilde{\mathbf{F}}_g, \widetilde{\mathbf{F}}_{\mathrm{loc}}\right), \qquad \Delta\mathbf{F} = \mathrm{Conv}_{\mathrm{out}}\left(\mathbf{F}_c\right). \tag{13}$$

The final SMFA output is obtained through residual aggregation:

$$\mathbf{F}_{\mathrm{out}} = \mathbf{F} + \Delta\mathbf{F}. \tag{14}$$

The residual connection retains the original information pathway and limits excessive modulation. SMFA thus combines variance-guided global modulation with lightweight local processing to recalibrate intermediate feature responses. As a feature-transformation module, however, it does not directly impose identity-level correspondence across high-level descriptor spaces, which motivates the ICL formulation introduced next.

3.3 Identity Conservation Learning

To address the remaining identity-semantic discrepancy across high-level descriptor spaces, we introduce Identity Conservation Learning (ICL), which promotes same-identity consistency while preserving inter-identity discrimination through prototype-level supervision.

As illustrated in Fig. 3, let the high-level fused descriptor groups defined in Section 3.1 be denoted by

$$\mathcal{Z} = \left\{ \mathbf{Z}^{(q)} \right\}_{q=1}^{Q}, \qquad \Omega_d \subseteq \{(a,b) \mid 1 \le a < b \le Q\}, \tag{15}$$

where Ω_d is a predefined set of descriptor-group pairs with valid identity correspondence. Group A and Group B in Fig. 3 represent an arbitrary selected pair $(a,b) \in \Omega_d$, rather than fixed modalities or branches.

For an arbitrary pair (a,b), assume that the current mini-batch contains K identities and P samples per identity in each descriptor group. The samples of identity k in group $g \in \{a,b\}$ are represented as

$$\mathbf{X}_k^{(g)} = \left[\mathbf{x}_{k,1}^{(g)}, \ldots, \mathbf{x}_{k,P}^{(g)} \right]^\top \in \mathbb{R}^{P \times D}, \qquad g \in \{a,b\}, \tag{16}$$

where D is the descriptor dimension and g indexes one descriptor space in the selected pair. Prototypes are constructed independently within each descriptor group rather than by first mixing samples from different modalities.

The sample features are L_2-normalized, and an intra-identity Gram matrix is computed as

$$\widetilde{\mathbf{x}}_{k,i}^{(g)} = \frac{\mathbf{x}_{k,i}^{(g)}}{\left\| \mathbf{x}_{k,i}^{(g)} \right\|_2}, \qquad \mathbf{G}_k^{(g)} = \widetilde{\mathbf{X}}_k^{(g)} \left(\widetilde{\mathbf{X}}_k^{(g)} \right)^\top. \tag{17}$$

The consistency scores and aggregation weights are then obtained by

$$\mathbf{s}_k^{(g)} = \mathbf{G}_k^{(g)} \mathbf{1}_P, \qquad \mathbf{w}_k^{(g)} = \operatorname{softmax}\left(\mathbf{s}_k^{(g)} \right), \tag{18}$$

where $\mathbf{1}_P$ is a P-dimensional all-ones vector. Samples that are more consistent with the remaining instances of the same identity receive relatively larger weights. The identity prototype and its normalized form are

$$\mathbf{p}_k^{(g)} = \left(\widetilde{\mathbf{X}}_k^{(g)} \right)^\top \mathbf{w}_k^{(g)}, \qquad \widehat{\mathbf{p}}_k^{(g)} = \frac{\mathbf{p}_k^{(g)}}{\left\| \mathbf{p}_k^{(g)} \right\|_2}. \tag{19}$$

This weighting scheme does not explicitly discard any sample; instead, it reduces the relative contribution of instances that are less consistent with the remaining samples of the same identity.

Although consistency-aware adaptive aggregation reduces the influence of noisy individual samples, the prototypes of the same identity may still remain displaced across paired descriptor spaces. For identity k, the prototype residual between groups a and b is defined as

$$\mathbf{r}_k^{(a,b)} = \widehat{\mathbf{p}}_k^{(a)} - \widehat{\mathbf{p}}_k^{(b)}. \tag{20}$$

The prototype residual consistency (PRC) loss is then formulated as

$$\mathcal{L}_{\mathrm{PRC}}(a,b) = \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{r}_k^{(a,b)}\right\|_2^2. \tag{21}$$

By penalizing the magnitude of the prototype residuals, this loss encourages same-identity prototypes from the paired spaces to approach a common semantic region. Because the constraint operates at the identity-prototype level, it is less sensitive to individual sample variations than instance-wise alignment. Nevertheless, PRC only reduces the discrepancy between same-identity prototypes and does not explicitly enforce discriminative separation among different identities.

To preserve such inter-identity boundaries, we further introduce bidirectional cross-group prototype classification (BPC). The normalized identity prototypes in each group are stacked as

$$\widehat{\mathbf{P}}^{(g)} = \left[\widehat{\mathbf{p}}_1^{(g)},\ldots,\widehat{\mathbf{p}}_K^{(g)}\right]^\top \in \mathbb{R}^{K\times D}, \qquad g\in\{a,b\}, \tag{22}$$

and the cross-group similarity matrix is defined by

$$\mathbf{S}^{(a,b)} = \widehat{\mathbf{P}}^{(a)}\left(\widehat{\mathbf{P}}^{(b)}\right)^\top. \tag{23}$$

The diagonal entries represent same-identity prototype matches, whereas the off-diagonal entries represent similarities between different identities. Prototype classification from group a to group b, and in the reverse direction, is formulated as

$$\mathcal{L}_{a\to b} = -\frac{1}{K}\sum_{k=1}^{K}\log\frac{\exp\left(S_{kk}^{(a,b)}\right)}{\sum_{j=1}^{K}\exp\left(S_{kj}^{(a,b)}\right)}, \tag{24a}$$

$$\mathcal{L}_{b\to a} = -\frac{1}{K}\sum_{k=1}^{K}\log\frac{\exp\left(S_{kk}^{(a,b)}\right)}{\sum_{i=1}^{K}\exp\left(S_{ik}^{(a,b)}\right)}. \tag{24b}$$

Their combination gives

$$\mathcal{L}_{\mathrm{BPC}}(a,b) = \mathcal{L}_{a\to b} + \mathcal{L}_{b\to a}. \tag{25}$$

This objective strengthens same-identity matching in both directions while treating other identity prototypes as negatives. The bidirectional formulation provides symmetric cross-group supervision without changing the underlying prototype-classification mechanism.

For each valid descriptor-group pair, the identity-conservation objective is defined as

$$\Psi\left(\widehat{\mathbf{P}}^{(a)},\widehat{\mathbf{P}}^{(b)}\right) = \mathcal{L}_{\mathrm{BPC}}(a,b) + \lambda_c\mathcal{L}_{\mathrm{PRC}}(a,b). \tag{26}$$

where λ_c balances the contribution of the PRC loss. In addition to direct descriptor-group pairs, the same pairwise objective is applied during training to predefined composite prototype-center pairs. Let $\mathcal{C}$ denote the composite centers aggregated from selected group-level prototypes and Ω_c denote their valid pair set. The complete ICL objective is

$$\mathcal{L}_{\mathrm{ICL}} = \sum_{(a,b)\in\Omega_d} \Psi\left(\widehat{\mathbf{P}}^{(a)}, \widehat{\mathbf{P}}^{(b)}\right) + \sum_{(u,v)\in\Omega_c} \Psi\left(\widehat{\mathbf{C}}^{(u)}, \widehat{\mathbf{C}}^{(v)}\right). \tag{27}$$

For a composite-center pair, Ψ is evaluated directly on the normalized composite prototype matrices. Through consistency-aware aggregation, prototype residual regularization, and bidirectional prototype classification, ICL jointly models same-identity semantic consistency and inter-identity discrimination across high-level descriptor spaces. ICL is used only during training and introduces no additional inference branch.

3.4 Overall Training Objective

HICR is optimized end-to-end using global and local identity classification losses, batch-hard triplet loss, and ICL. The global and local identity losses are applied to the corresponding global representation and local stripe representations, respectively. The batch-hard triplet loss is imposed on the 2048-dimensional pre-BN global GeM features and accumulated over the corresponding global feature groups, with a margin of $m_{\mathrm{tri}} = 0.3$. No additional triplet loss is applied to the local stripe descriptors.

The complete training objective is formulated as

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{glo}}\mathcal{L}_{\mathrm{id}}^{\mathrm{glo}} + \lambda_{\mathrm{loc}}\mathcal{L}_{\mathrm{id}}^{\mathrm{loc}} + \lambda_{\mathrm{tri}}\mathcal{L}_{\mathrm{tri}} + \lambda_{\mathrm{icl}}\mathcal{L}_{\mathrm{ICL}}, \tag{28}$$

where $\mathcal{L}_{\mathrm{id}}^{\mathrm{glo}}$ and $\mathcal{L}_{\mathrm{id}}^{\mathrm{loc}}$ denote the global and local identity classification losses, $\mathcal{L}_{\mathrm{tri}}$ denotes the accumulated batch-hard triplet loss, and $\mathcal{L}_{\mathrm{ICL}}$ is defined in Section 3.3. We set $\lambda_{\mathrm{glo}} = \lambda_{\mathrm{loc}} = \lambda_{\mathrm{tri}} = \lambda_{\mathrm{icl}} = 1.0$. The internal coefficient $\lambda_c = 1.3$ balances the PRC loss within ICL, whereas λ_{icl} controls the overall contribution of ICL to the complete objective.

The identity losses provide sample-level classification supervision, the triplet loss organizes the instance-level metric structure, and ICL further regularizes identity semantics at the prototype level. SMFA introduces no additional objective and is jointly optimized under Eq. (28). The supervision objectives are used only during training. During inference, only the normalized retrieval descriptor $\mathbf{f}_{i,\mathrm{ret}}^{m}$ is retained for cross-modal matching using cosine distance.

4 Experiments

This section evaluates HICR from three complementary perspectives: benchmark performance, controlled ablation, and representation analysis. We first describe the datasets, evaluation protocols, and implementation details, and then compare HICR with representative VI-ReID methods on SYSU-MM01, RegDB, and LLCM. Controlled experiments further examine component contributions, the stability of

individual component gains, internal module designs, parameter sensitivity, SMFA insertion layouts, and computational efficiency, followed by analyses of intermediate statistics, final embedding organization, and retrieval behavior.

4.1 Datasets and Evaluation Protocols

We evaluate HICR on three standard VI-ReID benchmarks: SYSU-MM01 [1], RegDB [2], and LLCM [12]. SYSU-MM01 is evaluated under the standard all-search and indoor-search settings, with each result averaged over 10 random gallery trials. For RegDB, we follow the standard protocol and report the average performance over 10 trials in both the VIS-to-IR and IR-to-VIS directions. LLCM is evaluated under its standard bidirectional protocol in the VIS-to-IR and IR-to-VIS directions. We report Rank-k, mean average precision (mAP), and mean inverse negative penalty (mINP), with Rank-1 and mAP used as the primary evaluation metrics.

4.2 Implementation Details

We adopt an ImageNet-pretrained ResNet-50 as the backbone and resize all input images to 288×144. The complete HICR configuration inserts five SMFA blocks at S2-1 and S2-2 in Stage 2 and at S3-1, S3-2, and S3-3 in Stage 3. Training augmentation includes random cropping after padding, horizontal flipping, random erasing with a probability of 0.5, and channel-adaptive grayscale augmentation. Each mini-batch contains eight identities, with four visible and four infrared images sampled for each identity, resulting in 64 images per batch.

Training follows a staged strategy. HICR is initialized from the fixed best checkpoint saved by the preceding pre-HICR training stage and is then fine-tuned for 20 epochs, corresponding to global epochs 100–119. We use stochastic gradient descent (SGD) with Nesterov momentum of 0.9 and a weight decay of 5×10^{-4}. The inherited high-level representation modules and the classification and projection heads are optimized with a learning rate of 10^{-3}, whereas the modality-specific stems, shared ResNet backbone, and SMFA blocks use 10^{-4}. Although the training script records the best intermediate epoch for monitoring, all reported results use the fixed epoch-119 checkpoint rather than an intermediate checkpoint selected according to test performance.

The global identity, local identity, and global batch-hard triplet losses are assigned weights of 1.0, and no additional triplet loss is applied to the local stripe descriptors. The coefficient of the PRC loss is set to $\lambda_c = 1.3$, the external ICL loss weight is set to $\lambda_{\text{icl}} = 1.0$, and the triplet margin is 0.3. During inference, only the L_2-normalized global-local fused retrieval descriptor is retained, and cross-modal matching is performed using cosine distance. The computational-efficiency protocol is detailed in Section 4.4.5.

4.3 Comparison with Representative VI-ReID Methods

We compare HICR with representative VI-ReID methods on SYSU-MM01, RegDB, and LLCM, using Rank-1 accuracy and mAP as the primary metrics. SYSU-MM01 is evaluated under the all-search and indoor-search settings, whereas RegDB and LLCM

Table 1: Comparison with representative VI-ReID methods on SYSU-MM01, RegDB, and LLCM. Rank-1 and mAP are reported in percentage (%). The best, second-best, and third-best results are marked in boldface, solid underlining, and italics, respectively. "–" denotes an unreported result. RegDB results are averaged over the standard 10 trials.

Methods	Venue	SYSU-MM01		RegDB		LLCM	
		All-search	Indoor-search	VIS-to-IR	IR-to-VIS	VIS-to-IR	IR-to-VIS
		Rank-1 / mAP	Rank-1 / mAP	Rank-1 / mAP	Rank-1 / mAP	Rank-1 / mAP	Rank-1 / mAP
MMN [37]	ACM MM 21	70.60 / 66.90	76.20 / 79.60	91.60 / 84.10	87.50 / 80.50	59.90 / 62.70	52.50 / 58.90
SFANet [24]	TNNLS 22	65.74 / 60.83	71.60 / 80.05	70.15 / 63.77	76.31 / 68.00	–	–
SPOT [5]	TIP 22	65.34 / 62.25	69.42 / 74.63	80.35 / 72.46	79.37 / 72.26	–	–
MAUM [30]	CVPR 22	71.68 / 68.79	76.52 / 81.94	87.87 / 85.09	86.95 / 84.34	–	–
G^2DA [6]	PR 23	63.94 / 60.73	71.06 / 76.01	73.95 / 65.49	69.67 / 61.98	–	–
PMT [38]	AAAI 23	67.53 / 64.98	71.66 / 76.52	84.83 / 76.55	84.16 / 75.13	–	–
DEEN [12]	CVPR 23	74.70 / 71.80	80.30 / 83.30	91.10 / 85.10	89.50 / 83.40	62.50 / 65.80	54.90 / 62.90
ACD [14]	TIFS 24	74.40 / 71.17	78.98 / 82.75	84.71 / 83.28	87.18 / 84.72	–	–
RLE [39]	NeurIPS 24	75.40 / 72.40	*84.70* / 87.00	92.80 / 88.60	91.00 / 86.60	–	–
HOS-Net [15]	AAAI 24	75.60 / *74.20*	84.20 / *86.70*	**94.70** / **90.40**	93.30 / 89.20	*64.90* / *67.90*	56.40 / *63.20*
AGCC [25]	PR 24	75.91 / 72.96	79.34 / 84.62	92.59 / 86.18	91.35 / 84.92	–	–
CycleTrans [28]	TNNLS 24	76.58 / 72.62	**87.22** / 84.92	91.50 / 87.00	90.60 / 85.60	–	–
CSCL [8]	TMM 24	75.72 / 72.80	80.80 / 83.56	92.16 / 84.27	89.66 / 85.07	–	–
$AGPI^2$ [35]	TIFS 25	72.23 / 70.58	83.45 / 84.25	89.03 / 83.89	87.91 / 83.04	61.78 / 65.08	*56.47* / 62.79
BDLF [23]	IJCAI 25	76.80 / 74.60	84.20 / 85.80	94.40 / 90.00	**94.50** / **89.60**	67.00 / 68.90	58.10 / 64.50
TSKD [40]	PR 26	76.60 / 73.00	82.70 / 85.30	91.10 / 81.70	89.90 / 80.50	–	–
WFLNet [29]	TOCE 26	*76.68* / 72.26	80.56 / 83.57	92.12 / 87.47	89.83 / 84.96	61.82 / 63.88	51.93 / 58.22
HICR (Ours)	–	**77.02** / **75.10**	85.52 / **87.54**	*94.17* / *88.88*	*92.56* / *86.94*	**68.48** / **70.66**	**60.15** / **66.57**

are evaluated in both the VIS-to-IR and IR-to-VIS directions. As reported in Table 1, HICR achieves seven best, one second-best, and four third-best results among the 12 reported metrics, ranking within the top three in every setting. This performance suggests that HICR is competitive across the evaluated benchmarks and retrieval directions.

On SYSU-MM01, HICR achieves 77.02% Rank-1 and 75.10% mAP under the all-search setting, slightly exceeding BDLF [23] by 0.22 and 0.50 percentage points, respectively. Under indoor-search, HICR obtains 85.52% Rank-1, which is 1.70 percentage points below CycleTrans [28], while achieving the best mAP of 87.54%, 0.54 percentage points above RLE [39]. Thus, although HICR does not obtain the highest top-ranked matching accuracy in the relatively controlled indoor setting, its mAP result suggests favorable overall ranking performance.

For RegDB, we follow the standard protocol and report the average performance over 10 trials. HICR obtains 94.17% Rank-1 and 88.88% mAP for VIS-to-IR retrieval and 92.56% Rank-1 and 86.94% mAP for IR-to-VIS retrieval, ranking third on all four metrics. The differences from the corresponding best results are 0.53, 1.52, 1.94, and 2.66 percentage points, respectively. Since several recent methods already achieve Rank-1 accuracy above 90% on RegDB, the relatively compressed performance range may partly limit the numerical differences among competing methods. Nevertheless, HICR remains within the top three on all four metrics.

On LLCM, HICR achieves the best results on all four metrics. Compared with BDLF, it improves Rank-1 and mAP by 1.48 and 1.76 percentage points in the VIS-to-IR direction and by 2.05 and 2.07 percentage points in the IR-to-VIS direction. These results indicate the effectiveness of HICR under the challenging low-light

Table 2: Component-level ablation on SYSU-MM01 under the all-search and indoor-search settings. The Non-local block is included as a generic relation-modeling reference. All configurations follow the same training and evaluation protocol. The best result in each column is highlighted in bold.

Setting	Components			All-search				Indoor-search			
	Non-local	SMFA	ICL	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Baseline	–	–	–	71.35	96.17	98.63	69.07	78.09	97.79	99.13	82.14
w/ Non-local	✓	–	–	70.65	95.94	98.64	68.20	77.32	97.20	98.87	81.25
w/ ICL	–	–	✓	73.97	96.97	99.01	71.92	82.49	98.66	99.55	85.71
w/ SMFA	–	✓	–	74.31	96.78	98.62	72.97	80.10	97.85	99.14	83.70
HICR (SMFA + ICL)	–	✓	✓	**77.02**	**97.40**	**99.27**	**75.10**	**85.52**	**98.94**	**99.62**	**87.54**

Table 3: Multi-seed stability analysis of the baseline and individual components on SYSU-MM01. Results are reported as mean ± standard deviation over five independent random seeds. The best mean in each setting is highlighted in bold. The complete HICR is not included in this experiment.

Method	Rank-1	Rank-10	Rank-20	mAP	mINP
All-search					
Baseline	71.52 ± 0.38	96.24 ± 0.16	98.69 ± 0.06	69.31 ± 0.40	56.96 ± 0.48
w/ SMFA	$\mathbf{74.31 \pm 0.33}$	96.84 ± 0.04	98.66 ± 0.07	$\mathbf{72.63 \pm 0.43}$	$\mathbf{61.41 \pm 0.68}$
w/ ICL	73.40 ± 0.42	$\mathbf{96.92 \pm 0.10}$	$\mathbf{99.19 \pm 0.11}$	71.45 ± 0.46	59.71 ± 0.54
Indoor-search					
Baseline	78.26 ± 0.50	97.74 ± 0.04	99.12 ± 0.04	82.23 ± 0.34	79.05 ± 0.33
w/ SMFA	79.79 ± 0.53	97.51 ± 0.06	98.92 ± 0.05	83.25 ± 0.36	80.09 ± 0.36
w/ ICL	$\mathbf{83.23 \pm 0.30}$	$\mathbf{99.21 \pm 0.15}$	$\mathbf{99.74 \pm 0.19}$	$\mathbf{86.20 \pm 0.26}$	$\mathbf{83.19 \pm 0.38}$

cross-modal setting of LLCM. Overall, Table 1 demonstrates the consistent competitiveness of HICR across multiple benchmarks and retrieval directions. The individual contributions and complementarity of SMFA and ICL are further examined in Section 4.4.

4.4 Ablation Studies

We conduct controlled experiments on SYSU-MM01 to examine the components and key configurations of HICR, including component contributions, multi-seed stability, internal module designs, parameter sensitivity, insertion layouts, and computational efficiency. Component-level and SMFA internal ablations are evaluated under both the all-search and indoor-search settings, while the remaining studies use all-search. ICL is disabled when analyzing SMFA, whereas SMFA is disabled in the independently trained ICL internal study. Parameter sensitivity is evaluated using the complete HICR. Rank-1 and mAP are used as the primary metrics. Unless otherwise specified, single-run ablations use seed 0, and SYSU-MM01 results are averaged over 10 random gallery trials. Table 3 reports five-seed statistics, whereas Table 5 is an independently retrained controlled experiment.

4.4.1 Component-Level Analysis

Table 2 compares the baseline, Non-local block, SMFA, ICL, and the complete HICR under the same protocol. Table 3 further reports five-seed results for the baseline, SMFA-only, and ICL-only configurations. Since the complete HICR is not included in the latter experiment, the multi-seed conclusions apply only to the individual components.

As shown in Table 2, SMFA improves Rank-1 and mAP by 2.96 and 3.90 percentage points under all-search and by 2.01 and 1.56 points under indoor-search. ICL yields corresponding gains of 2.62 and 2.85 points under all-search and 4.40 and 3.57 points under indoor-search. These results show that both components are independently effective, although their gains vary across evaluation settings.

Combining SMFA and ICL achieves the best result in every column. Compared with the baseline, HICR improves Rank-1 and mAP by 5.67 and 6.03 percentage points under all-search and by 7.43 and 5.40 points under indoor-search. Its consistent advantage over the two single-component variants supports the complementarity between intermediate feature-response calibration and high-level identity-semantic regularization. In contrast, the Non-local block does not improve the primary metrics or the overall retrieval performance under the current setting.

The five-seed results in Table 3 show that the gains of SMFA and ICL remain consistent across random initializations, with all standard deviations below 0.7 percentage points. Although minor fluctuations occur in the nearly saturated Rank-10 and Rank-20 metrics, both components consistently improve Rank-1, mAP, and mINP. Overall, the results support their independent effectiveness, joint complementarity, and reproducible single-component gains.

4.4.2 Internal Structure Analysis of SMFA

To examine the internal components of SMFA, we individually remove Global Statistical Modulation (GSM), the Local Detail Branch (LDB), and Variance-Aware Recalibration (VAR) from the complete module. ICL is disabled in all variants, and the remaining settings are kept unchanged.

As shown in Table 4, removing GSM causes the largest degradation. Rank-1 and mAP decrease by 3.23 and 4.38 percentage points under all-search and by 2.53 and 2.01 points under indoor-search, identifying GSM as the primary contributor within SMFA. Removing VAR also reduces Rank-1 and mAP by 2.72 and 3.59 points under all-search and by 2.03 and 1.64 points under indoor-search, supporting channel-wise variance as an informative modulation prior.

Removing LDB produces only minor and inconsistent changes across the reported metrics, indicating a limited independent contribution to aggregate retrieval performance. Overall, GSM provides the main performance gain, VAR supplies an important statistical prior, and LDB serves as a complementary local processing pathway. The corresponding layer-wise statistical changes are further examined in Section 4.5.1.

Table 4: Internal ablation of SMFA on SYSU-MM01. Each reduced variant removes one component from the complete SMFA. ICL is disabled in all experiments to isolate the effects of GSM, LDB, and VAR. The best result in each column is highlighted in bold.

Setting	SMFA components			All-search				Indoor-search			
	GSM	LDB	VAR	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
w/o GSM		✓	✓	71.08	96.16	98.84	68.59	77.57	**97.98**	**99.20**	81.69
w/o LDB	✓		✓	**74.38**	**96.88**	**98.86**	72.75	**80.58**	97.74	99.09	**83.81**
w/o VAR	✓	✓		71.59	96.33	98.70	69.38	78.07	97.66	99.11	82.06
Full SMFA	✓	✓	✓	74.31	96.78	98.62	**72.97**	80.10	97.85	99.14	83.70

Table 5: Internal structure analysis of ICL on SYSU-MM01 under the all-search setting. SMFA is disabled in all variants. This table reports an independently retrained controlled experiment using seed 0, with results averaged over 10 random gallery trials. Comparisons are made only among variants within this table and not against the component-level results in Table 2. The best result in each column is highlighted in bold.

Variant	Rank-1	Rank-10	Rank-20	mAP	mINP
No ICL	70.45	96.04	98.66	67.97	55.18
Uniform prototype	73.37	97.01	99.05	70.98	**58.51**
w/o PRC	72.16	96.89	99.01	70.26	58.09
One-way BPC	73.42	97.12	99.14	70.92	58.31
w/o BPC	73.18	96.78	98.93	70.25	57.32
Full ICL	**73.46**	**97.15**	**99.18**	**71.03**	58.49

4.4.3 Internal Structure Analysis of ICL

To examine the internal designs of ICL, we replace adaptive prototype construction with uniform averaging, remove the PRC loss, replace bidirectional prototype classification with a one-way formulation, and remove prototype classification entirely. SMFA is disabled in all variants. Since Table 5 reports an independently retrained controlled experiment, comparisons are restricted to variants within this table.

As shown in Table 5, Full ICL improves Rank-1, Rank-10, Rank-20, mAP, and mINP over No ICL by 3.01, 1.11, 0.52, 3.06, and 3.31 percentage points, respectively. Removing PRC decreases Rank-1 and mAP by 1.30 and 0.77 percentage points, respectively, while removing BPC reduces them by 0.28 and 0.78 points and causes a 1.17-point decrease in mINP. These results support PRC and BPC as the two principal, complementary constraints within ICL.

Replacing adaptive prototypes with uniform averaging or bidirectional BPC with its one-way counterpart produces only marginal changes. Full ICL achieves the best Rank-1, Rank-10, Rank-20, and mAP, whereas the uniform-prototype variant obtains a slightly higher mINP. Thus, adaptive prototype weighting and bidirectional classification mainly provide incremental refinements to the core ICL objective.

Table 6: Sensitivity analysis of the ICL coefficients on SYSU-MM01 under the all-search setting. The left group varies the internal PRC coefficient λ_c while fixing $\lambda_{\text{icl}} = 1.0$, and the right group varies the external ICL loss weight λ_{icl} while fixing $\lambda_c = 1.3$. The best result for each metric is highlighted in bold.

λ_c			λ_{icl}		
Value	Rank-1 (%)	mAP (%)	Value	Rank-1 (%)	mAP (%)
0.0	76.14	74.81	0.0	74.31	72.97
0.5	76.58	75.03	0.1	75.70	74.41
1.0	76.92	**75.13**	0.5	76.56	75.06
1.3	**77.02**	75.10	1.0	**77.02**	**75.10**
2.0	76.80	74.64	1.5	76.52	74.37
3.0	76.47	73.94	2.0	75.91	73.48

4.4.4 Parameter Sensitivity Analysis

We evaluate the sensitivity of ICL to the internal residual coefficient λ_c and the external loss weight λ_{icl} on SYSU-MM01 under the all-search setting. As reported in Table 6 and Fig. 4, λ_{icl} is fixed at 1.0 when varying λ_c, whereas λ_c is fixed at 1.3 when varying λ_{icl}.

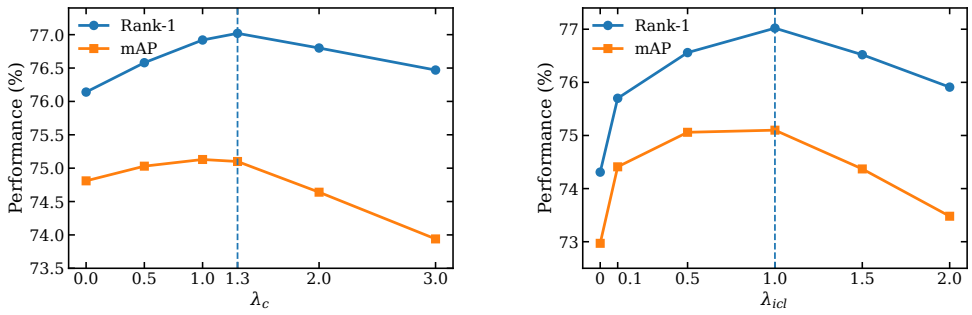
When $\lambda_c = 0$, only the PRC loss is disabled, while the remaining ICL supervision is retained. Increasing λ_c to a moderate value improves performance: mAP reaches its maximum of 75.13% at $\lambda_c = 1.0$, while Rank-1 reaches 77.02% at $\lambda_c = 1.3$, with a near-optimal mAP of 75.10%. Larger values reduce both metrics, suggesting that excessive PRC regularization may disturb the balance between semantic consistency and inter-identity discrimination.

Setting $\lambda_{\text{icl}} = 0$ removes ICL from the overall objective and corresponds to the SMFA-only configuration. Both Rank-1 and mAP reach their best values of 77.02% and 75.10% at $\lambda_{\text{icl}} = 1.0$, whereas larger weights degrade performance. The results also indicate that HICR is more sensitive to the overall ICL weight than to the internal residual coefficient over the tested ranges. We therefore use $\lambda_c = 1.3$ and $\lambda_{\text{icl}} = 1.0$ as the default HICR configuration.

4.4.5 Insertion Position and Computational Efficiency

To examine the influence of SMFA insertion positions, we construct five layouts across Stage 2 and Stage 3. S2-last and S3-last denote insertion only at S2-2 and S3-3, respectively; S2 and S3 cover all candidate positions in the corresponding stage; and S2+S3 uses all five positions. ICL is disabled in all variants, and the experiments are conducted on SYSU-MM01 under the all-search setting.

As shown in Table 7, the Stage-2-only layouts do not outperform the baseline, whereas both Stage-3 layouts provide clear improvements. S3 achieves 73.61% Rank-1, 71.68% mAP, and 60.15% mINP. The S2+S3 layout further reaches the best results



(a) Internal coefficient λ_c.

(b) External ICL loss weight λ_{icl}.

Fig. 4: Sensitivity of HICR to the ICL coefficients on SYSU-MM01 under the all-search setting: (a) the internal PRC coefficient λ_c, with $\lambda_{\text{icl}} = 1.0$; and (b) the external ICL loss weight λ_{icl}, with $\lambda_c = 1.3$.

Table 7: Influence of SMFA insertion positions on retrieval performance and computational efficiency under the SYSU-MM01 all-search setting. ICL is disabled in all variants to isolate the effect of the insertion layout. S2-last and S3-last denote insertion only at S2-2 and S3-3, respectively, whereas S2, S3, and S2+S3 cover all candidate positions in the corresponding stages. The best retrieval result in each column is highlighted in bold.

Layout	Rank-1 (%) ↑	mAP (%) ↑	mINP (%) ↑	Params (M) ↓	FLOPs (G) ↓	Latency (ms) ↓	Throughput (img/s) ↑
Baseline	71.35	69.07	57.03	132.46	18.30	8.81	1056.0
S2-last	70.57	68.81	56.63	132.92	18.81	9.18	994.3
S3-last	73.30	71.35	59.62	134.31	18.83	9.11	1005.7
S2	70.63	68.85	56.55	133.39	19.33	9.55	948.2
S3	73.61	71.68	60.15	138.00	19.88	10.08	931.2
S2+S3 (SMFA only)	**74.31**	**72.97**	**61.14**	138.93	20.92	10.96	850.9

Note: Retrieval results are obtained using seed 0 and averaged over 10 random gallery trials. Efficiency is measured in FP32 on a single NVIDIA RTX 3090 with an input size of 288 × 144. FLOPs are calculated as twice the number of MACs. Latency and throughput are measured with batch sizes of 1 and 64, respectively, using CUDA Events under `torch.inference_mode()`. We use 50 warm-up iterations followed by five measurement rounds of 200 forward passes each. Data loading and test-time flipping are excluded. ICL is training-only and therefore introduces no additional inference-time computation.

of 74.31%, 72.97%, and 61.14%, respectively. Although Stage-2-only insertion is ineffective under the current setting, the advantage of S2+S3 over S3 suggests that joint cross-stage modulation can provide complementary benefits.

Regarding the accuracy–efficiency trade-off, S3-last improves Rank-1 and mAP by 1.95 and 2.28 percentage points while increasing parameters, FLOPs, and latency by only 1.40%, 2.90%, and 3.41%, respectively. It therefore provides an efficiency-oriented configuration. By contrast, S2+S3 achieves the highest accuracy at a higher computational cost and is adopted as the default accuracy-oriented layout in HICR. The detailed efficiency measurement protocol is provided in Table 7.

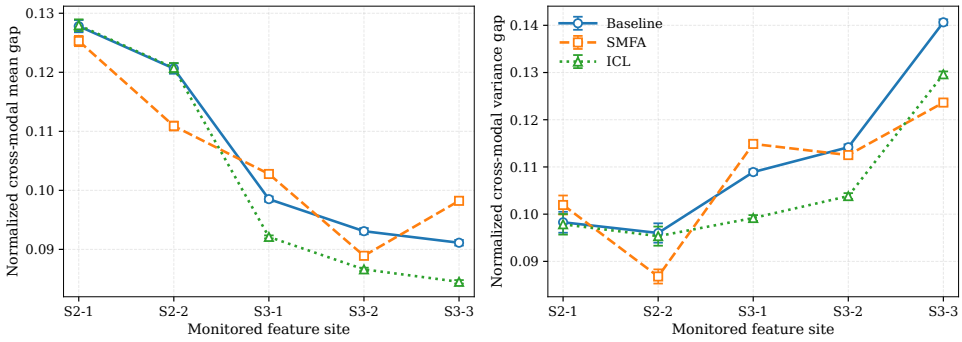


Fig. 5: Layer-wise cross-modal statistical discrepancies on SYSU-MM01. Normalized mean and variance gaps are compared across five monitored sites for the baseline, SMFA-only, and ICL-only variants. Lower values indicate smaller discrepancies.

4.5 Representation and Visualization Analysis

To further examine the learned representations, we provide quantitative and qualitative analyses from four complementary perspectives: layer-wise cross-modal statistical discrepancies, final cross-modal feature distributions, intra-/inter-class distance distributions, and retrieval rankings. The first two analyses characterize intermediate statistics and final embedding organization, whereas the latter two relate feature-space distances to retrieval behavior. These analyses complement the preceding retrieval and ablation results rather than serving as independent performance evidence.

4.5.1 Layer-wise Cross-Modal Statistical Discrepancy

We compare the normalized cross-modal mean and variance gaps of the baseline, SMFA-only, and ICL-only variants at five monitored sites in Stage 2 and Stage 3. The complete HICR is excluded to avoid confounding the individual effects of the two components. Fig. 5 reports the mean and standard deviation over five training seeds, where lower values indicate smaller statistical discrepancies between visible and infrared features.

As shown in Fig. 5, the effect of SMFA is site-dependent. It reduces both discrepancies at S2-2 and further decreases the corresponding gaps at several Stage-3 sites, but does not achieve the lowest value at every monitored position. In contrast, ICL has a weaker influence in Stage 2 and produces more consistent reductions in Stage 3. These observations are consistent with the respective roles of the two components: SMFA directly modulates intermediate feature responses, whereas high-level ICL supervision indirectly affects deeper representations through end-to-end optimization. The observed pattern also agrees with the insertion study in Section 4.4.5, where Stage 3 provides a more effective single-stage placement for SMFA.

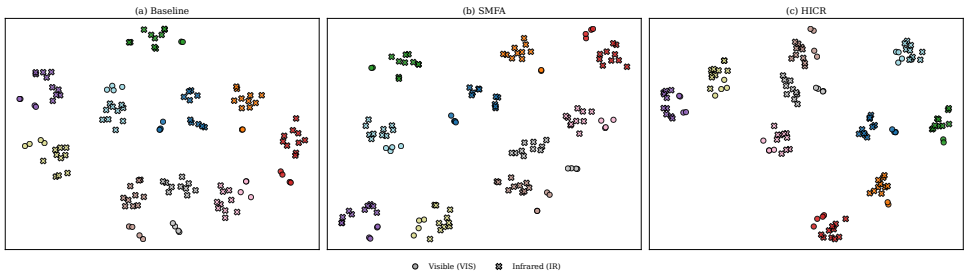


Fig. 6: t-SNE visualization of cross-modal feature distributions on SYSU-MM01: (a) baseline, (b) SMFA, and (c) HICR. The same color denotes the same identity, while circles and crosses represent visible and infrared samples, respectively. All panels use the same 136 samples from 10 identities.

Table 8: Quantitative feature-distribution analysis in the original L_2-normalized embedding space on SYSU-MM01. The same 136 samples from 10 identities as in Fig. 6 are used. Higher identity silhouette, modality silhouette closer to zero, and lower cross-modal prototype distance are preferred.

Method	Identity silhouette ↑	Modality silhouette → 0	Prototype distance ↓
Baseline	0.3129	0.0343	0.2584
w/ SMFA	0.3859	0.0313	0.2368
HICR	**0.4583**	**0.0071**	**0.1425**

4.5.2 Cross-Modal Feature Distribution Analysis

We visualize 136 visible and infrared samples from 10 SYSU-MM01 identities using t-SNE. The baseline, SMFA, and HICR use the same samples and dimensionality-reduction settings. As shown in Fig. 6, several identities in the baseline still form modality-separated subclusters. SMFA produces more compact identity clusters, while the complete HICR further improves cross-modal mixing within each identity and separation between different identities.

To avoid relying solely on the two-dimensional projection, Table 8 provides quantitative measurements in the original L_2-normalized embedding space before PCA and t-SNE. Identity silhouette measures identity-oriented clustering, modality silhouette values closer to zero indicate weaker modality-based grouping, and prototype distance measures the cosine distance between visible and infrared prototypes of the same identity.

Compared with the baseline, SMFA increases the identity silhouette score from 0.3129 to 0.3859 and reduces the prototype distance from 0.2584 to 0.2368. HICR further achieves an identity silhouette score of 0.4583, a modality silhouette score of 0.0071, and a prototype distance of 0.1425. These measurements indicate a clearer identity-oriented organization and stronger cross-modal consistency, consistent with the qualitative patterns observed in Fig. 6.

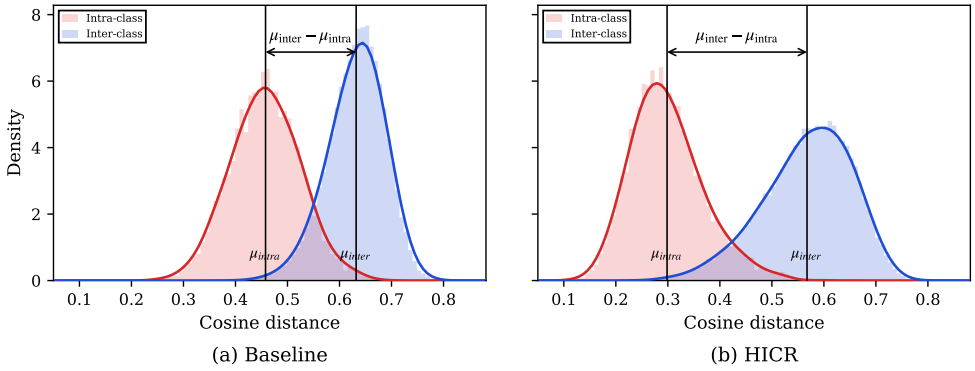


Fig. 7: Intra-class and inter-class cosine-distance distributions on SYSU-MM01 for (a) the baseline and (b) HICR. Both panels use the same cross-modal pair sets, histogram bins, and axis ranges. $\mu_{\text{inter}} - \mu_{\text{intra}}$ denotes the separation between the two distribution means.

4.5.3 Intra-/Inter-Class Distance Distributions

To examine the identity-discriminative structure of the final embedding space, we compute cross-modal cosine distances using the original L_2-normalized retrieval descriptors. Intra-class pairs consist of VIS–IR query-gallery samples sharing the same identity, whereas inter-class pairs contain different identities. The baseline and HICR use the same samples, complete pair sets, histogram bins, and axis ranges.

As shown in Fig. 7, HICR markedly shifts the intra-class distribution toward smaller distances and reduces its overlap with the inter-class distribution. The separation between the two distribution means, denoted by $\mu_{\text{inter}} - \mu_{\text{intra}}$, also becomes larger. Notably, this improvement is mainly associated with the substantial reduction in intra-class distances rather than an overall increase in the absolute inter-class distances. The result indicates improved cross-modal compactness within identities and stronger relative separability between intra- and inter-class pairs, consistent with the feature-distribution metrics in Table 8 and the retrieval improvements in Table 2.

4.5.4 Retrieval Ranking Results

Fig. 8 presents representative Top-10 retrieval results for the baseline and HICR on the same query examples from SYSU-MM01. Compared with the baseline, HICR places more correct matches at the top ranks and produces fewer high-ranked mismatches in the displayed cases. This ranking pattern is consistent with the improved intra-/inter-class distance distributions in Fig. 7 and the Rank-1 and mAP gains in Table 2, indicating that the improved feature-space organization translates into a more identity-consistent retrieval ordering. The visualization provides qualitative support for the quantitative results rather than independent performance evidence.

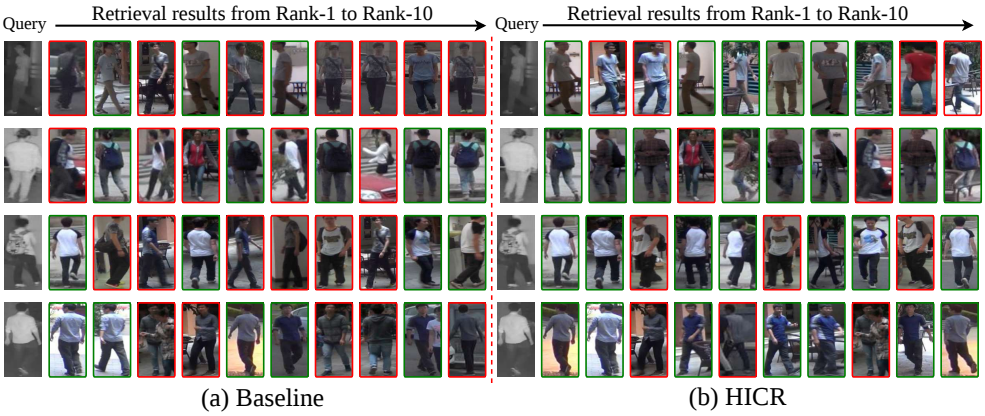


Fig. 8: Representative Top-10 retrieval results on SYSU-MM01 for the baseline and HICR. Green and red boxes denote correct and incorrect matches, respectively.

5 Conclusion

In this work, we presented HICR, a hierarchical identity-constrained representation learning framework with statistical calibration for visible–infrared person re-identification. HICR jointly addresses statistical perturbations in intermediate representations and identity-semantic shifts in high-level embeddings. SMFA recalibrates modality-sensitive intermediate responses, whereas ICL constrains identity semantics across multiple high-level descriptor spaces. Their joint optimization hierarchically connects intermediate statistical calibration with high-level identity-semantic regularization.

Experiments on SYSU-MM01, RegDB, and LLCM indicate that HICR is competitive across the evaluated benchmarks and retrieval directions. Controlled ablations support the effectiveness and complementarity of SMFA and ICL, while insertion-position, efficiency, and representation analyses further characterize the proposed design. Multi-seed evaluations indicate reproducible gains for the individual components. The five-position SMFA layout increases inference cost, whereas mini-batch prototype construction may be sensitive to batch composition. Future work will explore adaptive module deployment, memory-assisted prototypes, and generalization under more complex imaging conditions.

Funding. This work was supported by the National Natural Science Foundation of China (No. 62566013 and No. 62376232); the Excellent Youth Science and Technology Talent Program of Guizhou Province (QKHPTTRC-YQK[2023]028); the Leading Talent Workstation for Technological Innovation in Artificial Intelligence of Guizhou Province (No. QKHPT KXJZ[2025]022); the Guizhou Provincial Science and Technology Plan Project (No. QKHRC-KJFZ[2025]016); the Guizhou Province Academician

Innovation Team Workstation for Future Network (No. KXJZ[2024]008); and the Construction of the Data Intelligent Service Collaborative Innovation Center in Guiyang and Guian (QKH-ZYD[2025]006).

Author Contributions. All authors contributed to the study conception and design. Conceptualization: Rongyu Jia and Daoxun Xia; Methodology: Rongyu Jia; Software and experiments: Rongyu Jia; Data analysis: Rongyu Jia and Haiyang Chang; Writing–original draft preparation: Rongyu Jia; Writing–review and editing: Haiyang Chang and Daoxun Xia; Supervision: Daoxun Xia; Funding acquisition: Daoxun Xia. All authors read and approved the final manuscript.

Data availability and access. The datasets used in this study are publicly available benchmark datasets, including SYSU-MM01, RegDB, and LLCM. The experimental data generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethical and informed consent for data used. The datasets used in this manuscript are publicly available benchmark datasets for visible-infrared person re-identification. No new human participants were recruited, and no new private personal data were collected in this study. All datasets involved in the current study are listed in the Experiments section of the manuscript and were used only for academic research. All authors have approved the final manuscript and consented to its submission.

Competing Interests. The authors declare that they have no conflict of interest.

References

[1] Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380–5389 (2017)

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

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

[4] Xi, R., Huang, N., Lai, C., Zhang, Q., Han, J.: Fmcnet+: Feature-level modality compensation for visible-infrared person re-identification. IEEE Transactions on Neural Networks and Learning Systems **36**(7), 13247–13261 (2024)

[5] Chen, C., Ye, M., Qi, M., Wu, J., Jiang, J., Lin, C.-W.: Structure-aware positional transformer for visible-infrared person re-identification. IEEE Transactions on Image Processing **31**, 2352–2364 (2022)

[6] Wan, L., Sun, Z., Jing, Q., Chen, Y., Lu, L., Li, Z.: G2da: Geometry-guided dual-alignment learning for rgb-infrared person re-identification. Pattern Recognition **135**, 109150 (2023)

[7] Fang, X., Yang, Y., Fu, Y.: Visible-infrared person re-identification via semantic alignment and affinity inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11270–11279 (2023)

[8] Liu, M., Zhang, Z., Bian, Y., Wang, X., Sun, Y., Zhang, B., Wang, Y.: Cross-modality semantic consistency learning for visible-infrared person re-identification. IEEE Transactions on Multimedia **27**, 568–580 (2024)

[9] Ye, M., Lan, X., Wang, Z., Yuen, P.C.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE transactions on information forensics and security **15**, 407–419 (2019)

[10] Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X., Tao, D.: Hetero-center loss for cross-modality person re-identification. Neurocomputing **386**, 97–109 (2020)

[11] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence **44**(6), 2872–2893 (2021)

[12] Zhang, Y., Wang, H.: 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 (CVPR), pp. 2153–2162 (2023)

[13] Zhang, Y., Yan, Y., Lu, Y., Wang, H.: Adaptive middle modality alignment learning for visible-infrared person re-identification. International Journal of Computer Vision **133**(4), 2176–2196 (2025)

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

[15] Qiu, L., Chen, S., Yan, Y., Xue, J.-H., Wang, D.-H., Zhu, S.: High-order structure based middle-feature learning for visible-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 4596–4604 (2024)

[16] Feng, J., Wu, A., Zheng, W.-S.: Shape-erased feature learning for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22752–22761 (2023)

[17] Jiang, Y., Cheng, X., Yu, H., Liu, X., Chen, H., Zhao, G.: Dsaf: Dual space alignment framework for visible-infrared person re-identification. IEEE Transactions

on Multimedia (2025)

[18] Hua, X., Cheng, K., Lu, H., Tu, J., Wang, Y., Wang, S.: Mscmnet: Multi-scale semantic correlation mining for visible-infrared person re-identification. Pattern Recognition **159**, 111090 (2025)

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

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

[21] Huang, N., Xing, B., Zhang, Q., Han, J., Huang, J.: Co-segmentation assisted cross-modality person re-identification. Information Fusion **104**, 102194 (2024)

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

[23] Gong, Z., Wu, L., Xu, Y.: Base-detail feature learning framework for visible-infrared person re-identification. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pp. 1035–1043 (2025)

[24] Liu, H., Ma, S., Xia, D., Li, S.: Sfanet: A spectrum-aware feature augmentation network for visible-infrared person reidentification. IEEE Transactions on neural networks and learning systems **34**(4), 1958–1971 (2021)

[25] Yu, H., Cheng, X., Cheng, K.H.M., Peng, W., Yu, Z., Zhao, G.: Discovering attention-guided cross-modality correlation for visible–infrared person re-identification. Pattern Recognition **155**, 110643 (2024)

[26] Huang, Z., Liu, J., Li, L., Zheng, K., Zha, Z.-J.: Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 1034–1042 (2022)

[27] Wu, Q., Dai, P., Chen, J., Lin, C.-W., Wu, Y., Huang, F., Zhong, B., Ji, R.: Discover cross-modality nuances for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4330–4339 (2021)

[28] Wu, Q., Xia, J., Dai, P., Zhou, Y., Wu, Y., Ji, R.: Cycletrans: Learning neutral yet discriminative features via cycle construction for visible-infrared person

re-identification. IEEE Transactions on Neural Networks and Learning Systems **36**(3), 5469–5479 (2024)

[29] Yu, T., Cheng, D., Jiang, H., Chen, L., Qian, J., Kou, Q., Zhai, G.: Wavelet-based frequency feature learning for visible-infrared person re-identification. IEEE Transactions on Consumer Electronics (2026)

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

[31] Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Transactions on Information Forensics and Security **16**, 728–739 (2020)

[32] Li, Y., Zhang, T., Liu, X., Tian, Q., Zhang, Y., Wu, F.: Visible-infrared person re-identification with modality-specific memory network. IEEE Transactions on Image Processing **31**, 7165–7178 (2022)

[33] Li, H., Li, M., Peng, Q., Wang, S., Yu, H., Wang, Z.: Correlation-guided semantic consistency network for visible-infrared person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **34**(6), 4503–4515 (2023)

[34] Ye, M., Wu, Z., Chen, C., Du, B.: Channel augmentation for visible-infrared re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(4), 2299–2315 (2023)

[35] Alehdaghi, M., Josi, A., Cruz, R.M., Shamsolmoali, P., Granger, E.: Adaptive generation of privileged intermediate information for visible-infrared person re-identification. IEEE Transactions on Information Forensics and Security **20**, 3400–3413 (2025)

[36] Zhang, H., Cheng, S., Du, A.: Multi-stage auxiliary learning for visible-infrared person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **34**(11), 12032–12047 (2024)

[37] Zhang, Y., Yan, Y., Lu, Y., Wang, H.: Towards a unified middle modality learning for visible-infrared person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 788–796 (2021)

[38] Lu, H., Zou, X., Zhang, P.: Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1835–1843 (2023)

[39] Lei, T., Zhang, Y., Han, K., Dai, P., Zhang, Y., Wu, Y., Ji, R.: Rle: A unified

perspective of data augmentation for cross-spectral re-identification. Advances in Neural Information Processing Systems **37**, 126977–126996 (2024)

[40] Shi, J., Yin, X., Zhang, D., Zhang, Z., Xie, Y., Qu, Y.: Two-stage knowledge distillation for visible-infrared person re-identification. Pattern Recognition **169**, 111850 (2026)