Pattern Recognition 180 (2026) 114641

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Modality-Specific Prototypes for Disentangled Reconstruction in modality-missing object re-identification

Zhendong Xu [a,b], Aihua Zheng [a,b,*], Zi Wang [c], Chenglong Li [a], Dong Geng [b], Jin Tang [d]

[a] School of Artificial Intelligence, Anhui University, Hefei, 230601, China
[b] Anhui Provincial Key Laboratory of Intelligent Detection and Diagnosis for Traffic Infrastructure, Hefei, 230032, China
[c] School of Biomedical Engineering, Anhui Medical University, Hefei, 230032, China
[d] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China

ARTICLE INFO

*Keywords:*
Object re-identification
Multi-modal learning
Prototype learning
Modality-missing challenge

ABSTRACT

The common strategy for handling modality-missing in object re-identification (Re-ID) is to transform the available modalities into the missing one. However, direct feature-level or image-level transformation may suffer from the entanglement of modality-specific information and identity information, which leads to discrepancies in the reconstructed features in terms of both modality and identity. To address this issue, we propose the Modality-Specific Prototypes for Disentangled Reconstruction Network (PDRNet), which learns modality-specific priors and identity-discriminative residual representations while mitigating the adverse impact of information entanglement during the reconstruction process. This enables robust object Re-ID under missing-modality conditions. Specifically, we introduce modality prototypes to characterize modality-specific information and design an Adaptive Prototype-driven Disentanglement (APD) module. This module encourages the disentanglement between modality-specific components and identity-discriminative residual features through an orthogonality-based decorrelation constraint, thereby reducing their mutual interference. In addition, an entropy-distribution-based adaptive update strategy is employed to enable the prototypes to dynamically reflect modality statistics. Furthermore, we propose a Hierarchical Prototype Injection (HPI) module, which utilizes the missing-modality prototype in two stages during reconstruction. The prototype is first injected as a condition to guide the content reconstruction of identity representations, and then re-injected and refined through a feed forward network, achieving controlled and progressive modality-style injection while preserving identity discrimination and restoring modality consistency. Extensive experiments on multiple multi-modal object re-identification datasets demonstrate that the proposed PDRNet significantly outperforms existing methods under missing-modality scenarios. The code is publicly available at: https://github.com/skye-1201/PDRNet.

## 1. Introduction

With the advancement of machine learning techniques and the development of social intelligence, object Re-ID tasks have made significant progress [1,2]. Recently, multi-modal Object Re-ID tasks have received increasing attention due to the performance of the traditional single-modal Re-ID methods dropping dramatically when facing complex illumination and extreme weather. To utilize the information in paired multi-modal data [3–5], most methods extract various clues provided by heterogeneous modalities and interactively fuse them, thus improving the model's Re-ID performance [6,7]. Recent cross-modal Re-ID studies have also emphasized that reducing modality discrepancy and learning modality-invariant representations are crucial for robust retrieval across heterogeneous modalities [8,9]. In real-world applications, models often face the challenge of unpredictable modality-missing. Existing fusion/interaction-based methods may fail due to data absence, while shared representation learning methods are unable to leverage the advantages of modality-specific information [10,11].

To recover missing modality information, as illustrated in Fig. 1(a) and (b), existing methods typically reconstruct the missing modality by transforming the available one at either the image level or the feature level [12,13]. Specifically, image-level approaches rely on encoder–decoder architectures, while feature-level approaches employ attention mechanisms. Although these methods exploit information from the available modalities to compensate for the missing ones, they

* Corresponding author at: School of Artificial Intelligence, Anhui University, Hefei, 230601, China.
*E-mail addresses:* zhendongxu1201@foxmail.com (Z. Xu), ahzheng214@foxmail.com (A. Zheng), ziwang1121@foxmail.com (Z. Wang), lcl1314@foxmail.com (C. Li), fa923259253@outlook.com (D. Geng), tangjin@ahu.edu.cn (J. Tang).

https://doi.org/10.1016/j.patcog.2026.114641
Received 4 March 2026; Received in revised form 29 June 2026; Accepted 7 August 2026
Available online 14 August 2026

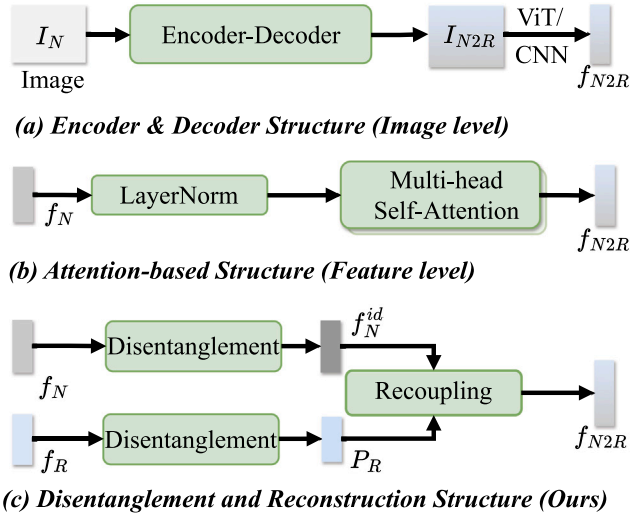



**Fig. 1.** Schematic diagram of different reconstruction strategies, using "NIR to RGB" as an example.

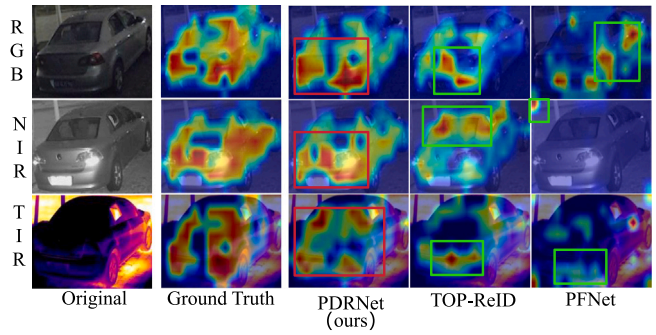



**Fig. 2.** The Grad-CAM [17] visualizations illustrate the reconstructed features under different modality-missing scenarios (from top to bottom: ground truth and reconstruction results under different missing modalities). The comparison highlights the reconstruction differences between the proposed method, PDRNet, and existing approaches, including TOP-ReID [12] and PFNet [3]. Regions highlighted in red indicate areas of attention in the reconstructed features.

often overlook the entanglement between modality-specific cues and identity-discriminative information. Such entanglement leads to reconstructed features that deviate from the true representations of the missing modality, thereby limiting the effectiveness of modality-missing Re-ID [11,14].

To address this issue, we propose a modality-specific prototypes for disentangled reconstruction method. As illustrated in Fig. 1(c), this framework encourages the disentanglement between modality-specific information and identity-discriminative information, enabling reconstructed representations to better approximate the true representations of the missing modality. PDRNet consists of two key components: An Adaptive Prototype-driven Disentanglement module that alleviates the entanglement between modality-specific cues and identity-discriminative information; and a Hierarchical Prototype Injection module that reconstructs robust missing-modality representations via two-stage prototype injection and progressive fusion.

Firstly, we disentangle modality-specific and identity information from the extracted object features. We adopt a prototype learning strategy, where modality-specific prototypes capture the general and shared characteristics within each modality. To obtain identity features that are less affected by modality-specific factors, we introduce an orthogonality-based decorrelation constraint to reduce the dependence between the two types of information. In contrast to prior methods that employ separate encoders for disentanglement [15,16], which increase memory overhead or rely on complex objectives, we design an Adaptive Prototype-driven Disentanglement (APD) module that reduces the dependence between modality-specific and identity-related information without additional encoders and updates modality prototypes adaptively. In APD, the identity-related feature is defined as the residual between a sample representation and its modality-specific prototype. We further promote disentanglement using a regularization loss that reduces the dependence between prototypes and identity features, while a cross-entropy loss enforces identity discriminability. Finally, we employ an entropy-based measure to weight prototype updates, down-weighting unreliable samples and enabling stable prototype refinement.

Secondly, to reconstruct discriminative representations of the missing modality from disentangled features, we design a Hierarchical Prototype Injection (HPI) module. Instead of one-shot modality transformation [3,12], HPI progressively injects the missing-modality prototype into the identity residual features extracted from the available modalities, which alleviates feature smoothing and pseudo-reconstruction issues commonly observed in direct transformation [18,19]. In Stage I, identity information from the available modalities is combined with the missing-modality prototype via feature-wise linear modulation, where the prototype serves as a conditioning signal to guide content reconstruction while preserving identity cues. In Stage II, the missing-modality prototype is explicitly re-injected and the fused features are refined through a feed forward network, introducing modality-specific information in a controlled manner to obtain the final reconstructed representation. This hierarchical design allows the prototype to influence reconstruction at different functional stages, balancing modality consistency and identity preservation. To further reduce cross-modality heterogeneity, we introduce a reconstruction consistency loss that aligns reconstructed features with their real counterparts in both distance and direction. As shown in Fig. 2, Grad-CAM [17] visualizations on RGB, NIR, and TIR reconstructions indicate that our approach suppresses modality-specific noise and yields more robust identity representations than existing methods.

We evaluate our framework under both missing modality and complete modality settings on four multi-modal object Re-ID datasets: RGBNT100 [20], WMVEID863 [5], MSVR310 [4], and RGBNT201 [3]. Extensive experimental results clearly demonstrate the effectiveness and generalization of our proposed approach.

The main contributions of this paper can be summarized as follows:

- We propose the disentangled reconstruction method (PDRNet), which addresses the reconstruction representation discrepancies caused by information entanglement in object Re-ID. PDRNet alleviates the entanglement between modality-specific information and identity information, and reconstructs more robust representations for the missing modality.
- We propose an Adaptive Prototype-driven Disentanglement (APD) module, which leverages modality-specific prototypes and orthogonality constraints to reduce the dependence between identity information and modality-specific information, and adaptively updates the prototypes guided by entropy.
- We propose a Hierarchical Prototype Injection (HPI) module that progressively injects missing-modality prototypes into identity residual features, enabling controlled and discriminative reconstruction under modality-missing scenarios.
- We perform extensive experimental comparisons with state-of-the-art methods on four multi-modal Re-ID datasets under both complete and modality-missing settings. These results fully demonstrate the effectiveness of the proposed PDRNet.

## 2. Related work

### 2.1. Multi-modal object re-ID

Recently, the robustness of infrared spectroscopy has attracted the attention of many researchers. Li et al. [20] propose the multi-modal vehicle datasets RGBN300 (RGB and NIR) and RGBNT100 (RGB, NIR, and TIR), along with a baseline method, HAMNet, for fusing information from different modalities. To overcome the differences between samples and modalities, Zheng et al. [4] propose CCNet to mitigate the discrepancies and a new multi-modal vehicle dataset, MSVR310. Zheng et al. [21] propose the IEEE framework to learn stable multi-modal person representations. Wang et al. [22] design a new training strategy, HTT, to fine-tune the model before inference and to enhance generalization. Zhang et al. [23] propose a new learning framework, EDITOR, for mitigating the effect of complexity in the Re-ID background. Yu et al. [24] propose a RSCNet that alleviates spectral heterogeneity in multi-spectral object re-identification via selective coupling with an Attention-Fourier Token Sparsification module and an Information Unification Constraint. Wang et al. [25] introduce a new framework, MambaPro, for multi-modal Re-ID to handle long sequential information across different modalities. Li et al. [26] enhance representation quality by exploiting learnable textual prompts to capture fine-grained information. In addition, Wang et al. [7] enrich semantic representations by integrating multi-modal textual cues. However, most of these methods target Re-ID under fully available modality scenarios. Their performance in incomplete modality settings suffers due to the common issue of missing multi-modal data in real-world applications.

### 2.2. Modality-missing object re-ID

With the development of multi-modal learning, various tasks in modality-missing scenarios attract increasing attention from researchers. These tasks can be roughly divided into two categories: (1) Generative methods, which replace missing data by generating similarly distributed data [27,28]; and (2) Joint learning approaches, which use a reconstruction module to supplement missing modalities by learning shared features across multiple modalities [29,30]. Recently, some works in the multi-modal Re-ID field have begun to consider modality-missing scenarios. Zheng et al. [3] propose PFNet, which transforms RGB into NIR and TIR to address missing-modality cases. To cope with different missing conditions, Wang et al. [12] propose TOP-ReID, which adapts to various missing-modality settings by inter-converting between the three modalities. Wang et al. [31] propose a multi-expert decoupled learning framework to alleviate the modality absence problem by learning more robust shared features. In addition, Zhang et al. [13] propose PromptMA, which addresses modality deficiency by leveraging prompt learning to enhance robust multi-modal complementary information. However, due to the entanglement of modality-specific and identity-related information within the learned representations, achieving a balance and clear separation between these two types of information in a unified transformation process remains challenging. This compromises the reconstruction quality and leads to representation discrepancies. Moreover, although shared information across modalities is jointly learned, relying solely on shared features during retrieval neglects the unique and discriminative modality-specific cues. In contrast, our proposed PDRNet mitigates reconstruction discrepancies caused by such entanglement. Additionally, modality-specific prototypes provide the missing discriminative cues for each reconstructed modality, further enhancing retrieval performance.

## 3. Method

In this section, we first introduce the baseline framework. Then, we present the proposed modality-specific prototype-based disentangled reconstruction method, with key components: (1) The Adaptive Prototype-driven Disentanglement (APD) module, which leverages prototypes to better disentangle identity-related information. (2) The Hierarchical Prototype Injection (HPI) module fuses the missing-modality prototype with the disentangled identity features from available modalities to reconstruct the missing-modality representation.

### 3.1. The baseline framework

As shown in Fig. 3, the proposed modality-missing object Re-ID framework extracts features from three modalities, namely visible (RGB), near-infrared (NIR), and thermal infrared (TIR), using visual encoder branches. The *input* can be denoted as:

$$input = [I_R, I_N, I_T], \tag{1}$$

where $[I_R, I_N, I_T]$ denotes the input tri-modality image. The size of each input image is $H \times W \times 3$. Due to the superior performance of Vision Transformer-based multi-modal Re-ID methods [7,31], we adopt a Vision Transformer (ViT) [32] as our backbone network to extract features, initialized with CLIP-pretrained weights to leverage large-scale vision–language knowledge for more robust representation learning. The features extracted by the CLIP-based backbone network can be expressed as follows:

$$f_m = \mathrm{CLIP}_m\left(I_m\right), \tag{2}$$

where $\mathrm{CLIP}_m$ denotes the ViT-based CLIP visual encoders, $I_m$ denotes input image and $m = [R, N, T]$. $f_R, f_N, f_T$ denote the features obtained through the CLIP-based feature extractor for the proposed APD and HPI, and the size of each modal feature is 768-dim.

### 3.2. Adaptive prototype-driven disentanglement

To model modality-specific characteristics, we employ a prototype learning strategy and maintain one modality prototype for each branch:

$$p = [p_R, p_N, p_T], \tag{3}$$

where $p_R$, $p_N$, and $p_T$ correspond to the RGB, NIR, and TIR branches, respectively. All prototypes are initialized as zero vectors with the same dimension as the corresponding modality class token.

**Adaptive prototype update.** Since different samples provide representations with varying reliability, we introduce a sample entropy-distribution-based weighting strategy to adaptively update modality prototypes. Given a batch feature $f_m^i \in \mathbb{R}^D$, we measure its uncertainty by entropy:

$$H(f_m^i) = -\sum_{d=1}^{D} \sigma(f_m^i)^d \log\left(\sigma(f_m^i)^d + \epsilon\right), \tag{4}$$

where $\sigma(\cdot)$ represents the softmax function, and $\epsilon$ is a small constant for numerical stability.

We convert uncertainty into a contribution weight via a softmin mapping and then normalize it within each mini-batch:

$$\alpha_m^i = \frac{\exp\left(-H(f_m^i)\right)}{\sum_{j=1}^{B} \exp\left(-H(f_m^j)\right)}, \tag{5}$$

where $B$ is the batch size. In this way, low-entropy samples are assigned larger weights for prototype refinement, while high-entropy samples are down-weighted to reduce unreliable updates.

The prototype is refined using a weighted aggregation strategy. During training, we update the modality prototype after each iteration as:

$$p_m \leftarrow (1 - \bar{\alpha}_m)\, p_m + \frac{1}{B}\sum_{i=1}^{B} \alpha_m^i f_m^i, \qquad \bar{\alpha}_m = \frac{1}{B}\sum_{i=1}^{B} \alpha_m^i, \tag{6}$$

where $m$ denotes the modality, $B$ is the batch size, and $i \in \{1, \ldots, B\}$ indexes samples in the current mini-batch. After being updated in this way, the modality prototype aggregates reliable modality-level

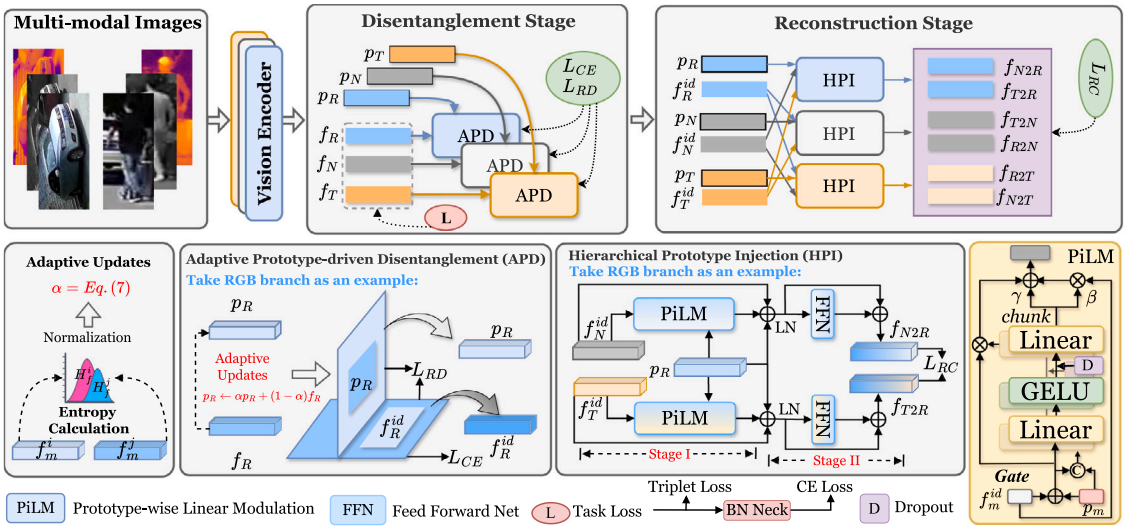



**Fig. 3.** Overview of the proposed PDRNet framework. The proposed framework extracts features from RGB, NIR, and TIR modalities using ViT-B/16 visual encoder branches. The prototype is introduced to learn the modality information. Adaptive Prototype-driven Disentanglement (APD) module is proposed to disentangle the identity information through the joint constraints of Representation Disentanglement ($\mathcal{L}_{RD}$) loss and cross-entropy loss. Meanwhile, the proposed Hierarchical Prototype Injection (HPI) module reconstructs the features using Reconstruction Consistency ($\mathcal{L}_{RC}$) loss constraints.

statistics and serves as the modality-specific prior for disentanglement in APD and missing-modality reconstruction in HPI.

**Warm-up.** To quickly bootstrap prototypes from the zero initialization, we apply a one-epoch warm-up by using a fixed update weight:

$$\alpha_m^i = \begin{cases} 0.99, & n \le n_w \\ \text{Eq. (5)}, & n > n_w, \end{cases} \tag{7}$$

where $n$ is the epoch index, $n_w = 1$.

**Prototype-driven disentanglement.** Disentangling modality-specific information and identity-discriminative information is important for better reconstruction of cross-modality features. We propose an Adaptive Prototype-driven Disentanglement module (APD) that utilizes prototypes to disentangle the identity information of each modality sample. Fig. 3 shows that the prototype feature $p_m$ is used to learn modality-specific information. The residual feature obtained by subtracting the prototype from the original feature is expected to preserve identity-discriminative information with reduced modality-specific interference. To reduce the dependency between the modality-specific prototype $p_m$ and the residual identity feature $f_m^{id}$, we introduce the following objective:

$$I(p_m; f_m^{id}) \to 0, \tag{8}$$

where $I(\cdot)$ denotes the mutual information. Inspired by this dependency-reduction objective and existing decorrelation-based redundancy-reduction studies [33,34], we adopt an orthogonality constraint as a tractable decorrelation regularization to reduce the linear correlation between $p_m$ and $f_m^{id}$, rather than strictly guaranteeing statistical independence. With normalized features, this constraint decreases the projection of the residual identity feature onto the modality-prototype direction. Therefore, we propose a Representation Disentanglement (RD) loss:

$$\mathcal{L}_{RD} = \frac{1}{B}\sum_{i=1}^{B}\left(p_m^{\top} f_m^{id}\right)^2, \tag{9}$$

where $p_m^T$ denotes the transpose of $p_m$. Further, to ensure that the residual identity feature $f_m^{id}$ remains identity-discriminative when the influence of generic modality information $P_m$ is reduced, we minimize the cross-entropy loss:

$$\mathcal{L}_{CE}^{id} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{s=1}^{S} y_{i,s}\log\left(g_s(f_m^{id,i})\right), \tag{10}$$

where $S$ denotes the total number of identities in the training set, $g(\cdot)$ denotes the classifier, $y_{i,s}$ denotes the ground truth of each sample.

Through joint training with Eqs. (9) and (10), APD encourages the residual identity-related feature $f_m^{id}$ to preserve identity-discriminative cues while reducing its linear dependence on the modality prototype $p_m$.

### 3.3. Hierarchical prototype injection

To generate missing-modality reconstructions with strong identity discriminability, we propose a Hierarchical Prototype Injection (HPI) module. HPI uses the prototype of the missing modality as a conditioning signal and injects it into the identity residual features disentangled from the available modalities in two stages: it first performs prototype-wise linear modulation on the identity content for robust reconstruction, and then re-injects the prototype to further refine the features, thereby restoring modality consistency while preserving identity-discriminative cues. To regularize the reconstruction process and avoid pseudo-reconstructions that are numerically close but semantically invalid, we further propose a reconstruction consistency loss $\mathcal{L}_{RC}$ for optimization.

**Stage I: prototype-wise linear modulation.** In the first stage, we propose a prototype-wise linear modulation scheme to robustly map the identity residual features of the available modality into the missing-modality space at the content level. The core idea is to use the missing-modality prototype $p_m$ as a conditioning signal to perform channel-wise scaling and shifting on the available-modality identity feature $f_a^{id,i}$, thereby injecting modality-specific statistics while preserving identity-discriminative structure for controllable reconstruction.

Specifically, we first add the missing-modality prototype to the available-modality identity feature and predict channel-wise scaling and shifting parameters:

$$[\gamma_m, \beta_m] = \phi(p_m + f_a^{id}), \qquad \gamma_m, \beta_m \in \mathbb{R}^D. \tag{11}$$

Here, $\phi(\cdot)$ is a one-layer MLP that outputs a $2D$-dimensional vector, which is then split into $\gamma_m$ and $\beta_m$ via chunk. $D$ denotes the feature dimension, and $i$ indexes samples.

To prevent large scaling factors from distorting the discriminative boundary of the reconstructed features, we stabilize the scale term as follows:

$$\gamma_m = 1 + \eta \cdot \tanh(\gamma_m), \qquad \eta \ll 1, \tag{12}$$

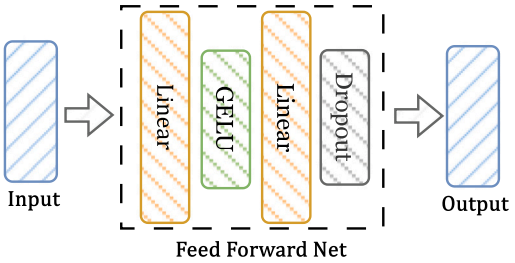



**Fig. 4.** The Feed Forward Network (FFN) consists of two linear layers, a GeLU activation, and dropout.

where $\tanh(\cdot)$ bounds the scaling magnitude and $\eta$ is a hyperparameter controlling the modulation strength, which improves stability in early training and during cross-modality transformation.

Moreover, linear modulation alone may be insufficient to capture nonlinear deviations in cross-modality mapping, potentially causing the loss of discriminative cues or injecting noise. Therefore, we further introduce an adaptive gated residual term to compensate for nonlinear errors and improve reconstruction reliability. Concretely, we compute channel-wise gating weights by:

$$g_m = \sigma(\phi_{gate}(p_m + f_a^{id})), \tag{13}$$

where $\phi_{gate}(\cdot)$ is a one-layer MLP and $\sigma(\cdot)$ is the sigmoid function, so that the gate suppresses unreliable residual injection in a channel-adaptive manner.

Next, we concatenate the missing-modality prototype with the available-modality identity feature and use an MLP to generate the residual correction term:

$$r_m = \phi_{res}\left([f_a^{id}, p_m]\right). \tag{14}$$

Here, $\phi_{res}(\cdot)$ consists of two linear layers, a GeLU activation, and dropout. $[\cdot,\cdot]$ denotes feature concatenation, and this branch explicitly models nonlinear corrections induced by cross-modality discrepancies.

Finally, the stage-I output is given by:

$$\mathrm{PiLM}(p_m, f_a^{id}) = \gamma_m \odot f_a^{id} + \beta_m \; + \; g_m \odot r_m, \tag{15}$$

where $\odot$ denotes channel-wise multiplication. Stage I injects modality information in a controllable manner via the prototype and enhances expressiveness through the gated residual, thereby improving the stability and consistency of missing-modality reconstruction while preserving identity discriminability.

**Stage II: Explicit re-injection.** In the second stage, we explicitly re-inject the missing-modality prototype $p_m$ to further strengthen modality-specific representations, and refine the fused output from Stage I to obtain the final reconstructed feature of the missing modality:

$$\begin{aligned} f_m^{Re} &= \mathrm{HPI}\left(p_m, f_a^{id}\right) \\ &= \mathrm{FFN}\left(\mathrm{PiLM}(p_m, f_a^{id}) + p_m + f_a^{id}\right), \end{aligned} \tag{16}$$

where $\mathrm{FFN}(\cdot)$ denotes the feed forward network shown in Fig. 4. By adding $p_m$ to the Stage I output, we perform a second prototype injection, which supplements missing-modality specific information while keeping the content representation stable, thereby improving modality consistency of the reconstructed features.

It is worth emphasizing that when only one modality is missing, each available modality can independently generate its corresponding reconstruction. Taking the case with two available modalities as an example, we obtain two reconstructed features $f_m^{Re1}$ and $f_m^{Re2}$, and the final reconstruction is produced by simple averaging:

$$f_m^{Re} = \frac{f_m^{Re1} + f_m^{Re2}}{2}. \tag{17}$$

**Reconstruction consistency loss.** In order to obtain robust reconstructed features, most existing methods approximate the distance between reconstructed and real features by $\mathcal{L}_{mse}$. However, relying solely on $\mathcal{L}_{mse}$ may potentially generate numerically close but semantically invalid reconstructed features, especially when there is heterogeneity between multiple modalities. Since the identity information of the same sample is consistent across modalities, we propose a reconstruction consistency loss ($\mathcal{L}_{RC}$) to optimize reconstructed features in terms of both feature distance and orientation. As an example, the $\mathcal{L}_{RC}$ for reconstructing RGB features is:

$$\mathcal{L}_{RC} = \frac{1}{B}\sum_{i=1}^{B}\left[\frac{1}{D}\left\|f_m^i - f_m^{Re,i}\right\|_2^2 - \frac{f_m^{Re,i}\cdot f_m^i}{\|f_m^{Re,i}\|\|f_m^i\|} + 1\right]. \tag{18}$$

The MSE term is normalized by the feature dimension $D$ to alleviate scale imbalance with the bounded cosine similarity term. The reconstruction consistency loss constrains semantic consistency between reconstructed and real features by enforcing both proximity and directional alignment, which avoids semantic distortion and improves the robustness and generalization ability of the reconstructed features. Moreover, by encouraging reconstructed features to align with real features, $\mathcal{L}_{RC}$ indirectly reduces inter-modality heterogeneity and benefits both missing-modality and full-modality settings.

### 3.4. Loss functions

We optimize the proposed framework during training with cross-entropy loss [35], triplet loss [36], the proposed Representation Disentanglement loss ($\mathcal{L}_{RD}$), and Reconstruction Consistency loss ($\mathcal{L}_{RC}$). The CE loss encourages the network to distinguish different identities, while the triplet loss enlarges the distance between negative pairs and reduces the distance between positive pairs. Therefore, $\mathcal{L}_{CE}$ and $\mathcal{L}_{tri}$ serve as the Re-ID task-related losses for supervising the backbone network.

$$\mathcal{L}_{task} = \mathcal{L}_{CE} + \mathcal{L}_{tri}. \tag{19}$$

Finally, the total training loss of the proposed framework is given as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_1\mathcal{L}_{RD} + \lambda_2\mathcal{L}_{RC}. \tag{20}$$

Here, $\lambda_1$ and $\lambda_2$ are two tunable hyperparameters that control the weights of the loss terms associated with the two proposed modules, so as to maintain a balanced collaboration between them.

The pseudo-code of the proposed PDRNet is presented in Algorithm 1.

**Algorithm 1** The algorithm for PDRNet

**Require:** Input feature $f_m$; Prototype $p_m$; Batch size $B$
**Ensure:** Reconstructed feature $\bar{f}$
**for** $i = 1$ to *epochs* **do**
  `# Prototype update process`
  Initialize three modality prototypes $p$ and update each prototype $p$ via Eq. (6)
  **for** $j = 1$ to $B$ **do**
  `# APD`
    Disentangle the identity information $f_m^{id,j}$ using $p$ via Eq. (9)
  `# HPI`
    Reconstruct $\bar{f}_m$ via Eq. (16) and Eq. (17)
  **end for**
  `# Loss Function`
  Update the feature $F$ and $\bar{f}$ with Eq. (20)
  `# Optimizer`
  Update the encoder weights using the Adam optimizer
**end for**

## 4. Experiments

### 4.1. Experimental data and evaluation metrics

**Datasets.** We evaluate our proposed framework on four benchmark datasets for multi-modal object Re-ID: RGBNT201 [3], MSVR310 [4], RGBNT100 [20], and WMVEID863 [5].

RGBNT201 [3] is the first multi-modal person Re-ID dataset, containing 4787 RGB–NIR–TIR image triplets from 201 identities. It is split into 141 identities for training, 30 for validation, and 30 for testing.

MSVR310 [4] dataset comprises 2087 image triplets (RGB, NIR, TIR) from 310 vehicle identities. The training set contains 1032 samples from 155 identities, while the remaining 155 identities contribute 1055 images for the gallery. A subset of 591 samples from 52 identities is randomly selected as the query set.

RGBNT100 [20] dataset consists of 17250 image triplets from 100 vehicle identities, equally split between training and testing. Specifically, 8675 images from 50 identities are used for training, and 8575 images from the remaining 50 identities are used for testing, with 1715 samples designated as queries. The RGB and NIR modalities are captured at 1920 × 1080 resolution, whereas TIR images are at 640 × 480.

WMVEID863 [5] dataset includes 4709 image triplets from 863 vehicle identities across 8 camera views. It utilizes 575 identities (3314 samples) for training and 288 identities (1309 samples) for testing. The gallery set comprises all test samples, while the query set includes 928 samples from 205 randomly selected test identities. Notably, WMVEID863 [5] is the first multi-spectral vehicle Re-ID benchmark designed to address the challenges posed by intense illumination conditions.

**Evaluation Metrics.** In our experiments, we evaluate our Re-ID model employing the industry-standard Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) metrics. CMC provides a probabilistic measure of match accuracy within top-k results, reporting Rank-1 (R-1), Rank-5 (R-5), and Rank-10 (R-10) values. The mAP score calculates overall precision and recall from all queries.

### 4.2. Missing setting and implementation details

**Modality-Missing Scenarios Setting:** (1) Fixed Missing: For input pairs of test samples, one or two modalities are fixed missing. (2) Random Missing: For the input test sample pairs, the samples are randomly missing at a certain ratio based on the random missing setting in CCNet [4], which indicates the probability of missing samples in random partial modality.

**Implementation Details:** We implement our method using PyTorch on one GeForce RTX 4090 GPU. We adopt CLIP-pretrained visual encoders as the backbone architecture. In the comparison tables, the symbol † denotes CLIP-based methods, ∗ denotes ViT-based methods, and the remaining methods are CNN-based. All methods are compared under the same dataset splits, modality-missing settings, and evaluation metrics. Nevertheless, comparisons between CNN-based methods and CLIP-pretrained ViT-based methods are not fully backbone-controlled due to differences in backbone architectures and pretraining strategies. For the RGBNT201 [3] dataset, we conduct experiments with a shared-parameter visual encoder and resize the input images to 256 × 128 pixels. For multi-modal vehicle Re-ID datasets, we use modality-specific visual encoders with independent parameters and resize the input images to 128 × 256 pixels. We apply standard training augmentations, including random cropping and horizontal flipping. In particular, during inference under the complete-modality setting, we only use the original features extracted from the backbone for retrieval, without executing the HPI module for missing-modality reconstruction. The network is optimized using Adam with an initial learning rate of $3.5 \times 10^{-4}$; for MSVR310, we use a smaller learning rate of $5 \times 10^{-6}$. All models are trained for 80 epochs. We set the loss weights to $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$. Unless otherwise specified, "M (·)" stands for missing modality. The best and second-best results are highlighted in **bold** and underlined, respectively. For the t-SNE visualization experiments, we set the test batch size to 64 and fix the random seed to 1111. The same setting is used for all compared visualization variants.

### 4.3. Comparison with state-of-the-art methods

**Evaluation on Modality-Missing Scenarios.**

**Fixed Missing:** As shown in Table 1, we compare PDRNet with existing methods for the fixed modality-missing problem in multi-modal Re-ID. On the multi-modal object Re-ID datasets, PDRNet shows strong performance across various modality-missing scenarios. On the multi-modal person Re-ID dataset, although PDRNet does not achieve state-of-the-art performance in every missing setting, it obtains the best average mAP and Rank-1, demonstrating more balanced robustness under different fixed modality-missing conditions. On the three multi-modal vehicle Re-ID datasets, PDRNet achieves the best performance on most evaluation metrics. On multiple vehicle Re-ID datasets, our method brings clear performance improvements. Under extreme modality-missing conditions, PDRNet shows a pronounced advantage over existing methods, demonstrating its robustness to various missing-modality cases. In addition, the missing-modality comparisons across datasets indicate that different modalities behave differently on different datasets. For example, on the strong-illumination multi-modal vehicle dataset WMVEID863 [5], RGB and NIR modalities are more easily affected by strong illumination, while TIR is less sensitive to illumination changes. Consequently, when TIR is missing, the performance drop is larger than that on other vehicle datasets, indicating that inter-modality imbalance can directly affect the final performance.

**Random Missing:** As shown in Fig. 5, we further analyze the performance of PDRNet under random ratio modality-missing settings. Specifically, during testing, we randomly drop different modality samples within the current mini-batch. To ensure a fair and reliable evaluation, for each missing ratio, we report the average performance over 10 independent runs as the final result at that ratio. We compare PDRNet with two state-of-the-art methods, DeMo [31] and PromptMA [13], under the same random missing protocol. The results show that PDRNet significantly outperforms existing methods. As the missing ratio increases, PDRNet maintains stronger robustness under more severe modality-missing conditions. Overall, the comparisons across different random modality-missing scenarios demonstrate the stability and superiority of our method for the random missing task.

**Evaluation on Complete Modality Scenarios.** In Table 2, we further compare PDRNet with state-of-the-art methods under the complete-modality setting. The results show that, although PDRNet is primarily designed for the modality-missing task, it still achieves the best or second-best performance on all four datasets, indicating strong generalization and robustness. Compared with fusion and feature extraction strategies tailored only for complete modalities, PDRNet not only learns robust reconstructions for missing modalities, but also enhances fine-grained identity feature extraction via adaptive disentanglement (APD). Moreover, the missing-modality reconstruction (HPI) encourages cross-modality feature alignment in a shared space, effectively alleviating performance degradation caused by modality heterogeneity.

### 4.4. Ablation study

**Effects of Key Modules.** To validate the effectiveness of the proposed modules under different scenarios, we conduct ablation studies under both complete-modality and fixed modality-missing settings. As shown in Table 3, Adaptive Prototype-driven Disentanglement (APD) improves the average robustness by disentangling modality-specific and identity-related information. However, APD does not improve all metrics in every missing case, since it mainly enhances feature disentanglement but does not directly recover missing-modality information.

**Table 1**
Experimental results of fixed modality-missing tasks on RGBNT201, MSVR310, RGBNT100 and WMVEID863. In the comparison tables, the symbol † denotes CLIP-based methods, ∗ denotes ViT-based methods, and the remaining methods are CNN-based.

| RGBNT201 | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | M (RGB) | | M (NIR) | | M (TIR) | | M (RGB+NIR) | | M (RGB+TIR) | | M (NIR+TIR) | | Average | |
| | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 |
| PFNet [3] | – | – | 31.9 | 29.8 | 25.5 | 25.8 | – | – | – | – | 26.4 | 23.4 | – | – |
| TOP-ReID* [12] | 54.4 | 57.5 | 64.3 | 67.6 | 51.9 | 54.5 | 35.3 | 35.4 | 26.2 | 26.0 | 34.1 | 31.7 | 44.4 | 45.5 |
| DESANet* [37] | 54.1 | 55.1 | 65.0 | 67.3 | 56.6 | 56.6 | 35.7 | 38.2 | 28.5 | 28.8 | 34.2 | 29.8 | 45.7 | 46.0 |
| MFRNet† [38] | 64.7 | 65.2 | 72.3 | 76.1 | 51.6 | 49.5 | 41.4 | 43.4 | 27.3 | 27.9 | 37.2 | 35.6 | 49.1 | 49.6 |
| DeMo† [31] | 63.3 | 65.3 | 72.6 | 75.7 | 56.2 | 54.1 | 45.6 | 46.5 | 26.3 | 24.9 | 40.3 | 38.5 | 50.7 | 50.8 |
| PromptMA† [13] | 67.4 | 68.4 | 72.5 | 75.7 | 58.9 | 57.3 | 51.5 | 53.0 | 30.4 | **33.3** | **43.9** | 42.2 | 54.1 | 55.0 |
| MDReID† [39] | 67.0 | 67.9 | **75.8** | **80.9** | 61.7 | 59.8 | 48.6 | 51.1 | 31.4 | 29.2 | 42.1 | 40.0 | 54.4 | 54.8 |
| Miss-ReID† [40] | 66.6 | 68.2 | 72.4 | 75.5 | **63.2** | **63.8** | 47.2 | 49.5 | **34.5** | **33.3** | **43.9** | **44.3** | 54.6 | 55.7 |
| PDRNet† (Ours) | **68.8** | **69.1** | **75.8** | 79.8 | 59.7 | 58.4 | **51.7** | **53.1** | 30.4 | 30.7 | 43.3 | 43.9 | **55.0** | **55.8** |
| RGBNT100 | | | | | | | | | | | | | | |
| CCNet [4] | 66.8 | 90.2 | 73.2 | 92.2 | 60.0 | 82.9 | 44.4 | 75.0 | 42.4 | 63.8 | 49.5 | 69.8 | 56.1 | 79.0 |
| TOP-ReID* [12] | 70.6 | 90.6 | 77.9 | 94.5 | 64.0 | 81.5 | 42.5 | 69.3 | 45.9 | 65.4 | 55.4 | 77.8 | 59.4 | 80.0 |
| Miss-ReID† [40] | 77.3 | 94.9 | 81.9 | 96.3 | 70.3 | **89.7** | 47.1 | 78.0 | 57.7 | 78.2 | 65.6 | **86.8** | 66.6 | 87.3 |
| PromptMA† [13] | 80.3 | **95.8** | 83.5 | 96.8 | 69.3 | 85.8 | 48.0 | 74.3 | 56.2 | 75.2 | 64.2 | 82.0 | 66.9 | 85.0 |
| DeMo† [31] | 81.0 | 94.5 | 84.1 | 96.5 | 71.1 | 87.6 | 50.2 | 73.7 | 59.6 | 78.1 | 66.3 | 82.8 | 68.7 | 85.5 |
| PDRNet† (Ours) | **82.7** | 95.0 | **88.0** | **98.2** | **75.1** | 88.6 | **53.8** | **79.9** | **61.3** | **80.3** | **67.3** | 84.0 | **71.4** | **87.7** |
| MSVR310 | | | | | | | | | | | | | | |
| CCNet [4] | 25.5 | 43.7 | 27.2 | 42.3 | 26.4 | 41.6 | 10.8 | 30.1 | 20.6 | 34.2 | 24.1 | 37.2 | 22.4 | 38.2 |
| TOP-ReID* [12] | 23.5 | 41.1 | 28.6 | 41.8 | 31.1 | 45.9 | 10.8 | 23.5 | 22.3 | 40.6 | 26.5 | 36.5 | 23.8 | 38.2 |
| DeMo† [31] | 36.9 | 55.3 | 43.1 | 56.5 | 46.1 | 60.9 | 10.5 | 24.2 | 34.1 | 53.5 | 40.8 | 53.6 | 35.3 | 50.7 |
| PDRNet† (Ours) | **49.0** | **68.7** | **55.0** | **73.6** | **53.7** | **71.1** | **25.7** | **40.6** | **40.2** | **56.0** | **49.5** | **65.1** | **44.6** | **61.6** |
| WMVEID863 | | | | | | | | | | | | | | |
| CCNet [4] | 58.7 | 64.8 | 59.8 | 64.9 | 53.5 | 58.7 | 51.9 | 56.8 | 43.0 | 44.1 | 46.7 | 48.6 | 52.3 | 56.3 |
| TOP-ReID* [12] | 63.2 | 69.9 | 66.8 | 73.7 | 55.3 | 59.8 | 60.3 | 65.7 | 46.8 | 49.5 | 53.0 | 56.4 | 57.6 | 62.5 |
| DeMo† [31] | 66.5 | 73.3 | 65.9 | 73.3 | 55.5 | 59.8 | 58.4 | 63.9 | 49.7 | 52.1 | 50.4 | 52.8 | 57.7 | 62.5 |
| PDRNet† (Ours) | **69.5** | **76.2** | **70.6** | **76.3** | **61.9** | **67.3** | **63.7** | **70.1** | **55.1** | **59.4** | **59.7** | **63.5** | **63.4** | **68.8** |

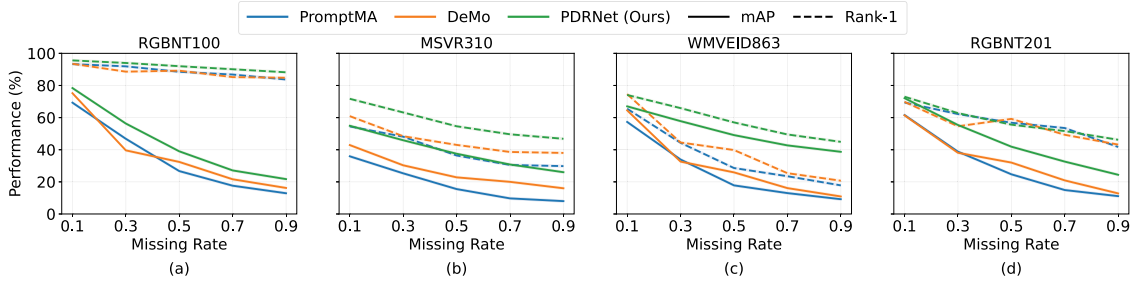



**Fig. 5.** Experimental results under percentage-based random modality-missing conditions on RGBNT100, MSVR310, WMVEID863, and RGBNT201.

**Table 2**
Comparison with state-of-the-art methods on RGBNT201, MSVR310, RGBNT100 and WMVEID863 datasets. In the comparison tables, the symbol † denotes CLIP-based methods, ∗ denotes ViT-based methods, and the remaining methods are CNN-based.

| Methods | Publication | RGBNT201 | | | | MSVR310 | | RGBNT100 | | WMVEID863 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mAP | R-1 | R-5 | R-10 | mAP | R-1 | mAP | R-1 | mAP | R-1 | R-5 | R-10 |
| CCNet [4] | INFFUS 2023 | – | – | – | – | 36.4 | 55.2 | 77.2 | 96.3 | 50.3 | 52.7 | 69.6 | 75.1 |
| TOP-ReID* [12] | AAAI 2024 | 72.3 | 76.6 | 84.7 | 89.4 | 35.9 | 44.6 | 81.2 | 96.4 | 67.7 | 75.3 | 80.8 | 83.5 |
| FACENet* [5] | INFFUS 2025 | – | – | – | – | 36.2 | 54.1 | 81.5 | 96.9 | 69.8 | 77.0 | 81.0 | 84.2 |
| EDITOR* [23] | CVPR 2024 | 66.5 | 68.3 | 81.1 | 88.2 | 39.0 | 49.3 | 82.1 | 96.4 | 65.6 | 73.8 | 80.0 | 82.3 |
| ICPL-ReID† [26] | TMM 2025 | 75.1 | 77.4 | 84.2 | 87.9 | 56.9 | **77.7** | 87.0 | **98.6** | 67.2 | 74.0 | 81.3 | 85.6 |
| MambaPro† [25] | AAAI 2025 | 78.9 | 83.4 | 89.8 | 91.9 | 47.0 | 56.5 | 83.9 | 94.7 | 69.5 | 76.9 | 80.6 | 83.8 |
| PromptMA† [13] | TIP 2025 | 78.4 | 80.9 | 87.0 | 88.9 | 55.2 | 64.5 | 85.3 | 97.4 | – | – | – | – |
| DeMo† [31] | AAAI 2025 | 79.0 | 82.3 | 88.8 | 92.0 | 49.2 | 59.8 | 86.2 | 97.6 | 68.8 | 77.2 | 81.5 | 83.8 |
| IDEA† [7] | CVPR 2025 | 80.2 | 82.1 | 90.0 | 93.3 | 47.0 | 62.4 | 87.2 | 96.5 | – | – | – | – |
| MFRNet† [38] | ICML 2025 | 80.7 | **83.6** | 91.9 | **94.1** | 50.6 | 64.8 | 88.2 | 97.4 | – | – | – | – |
| PDRNet† | Ours | **81.4** | 83.0 | **92.3** | 93.9 | **59.0** | 76.0 | **89.7** | 98.3 | **71.8** | **77.8** | **85.6** | **88.4** |

For example, APD brings limited gains or slight degradation in several specific settings, indicating that disentanglement alone is not sufficient to fully address modality-missing feature recovery. By further introducing Hierarchical Prototype Injection (HPI), the complete PDRNet model achieves the best average performance under fixed modality-missing settings on both datasets, demonstrating stronger robustness to missing modalities. It is worth noting that HPI does not dominate every individual metric. On WMVEID863 [5], the gains in several settings are

**Table 3**
Ablation experiments with different combinations of complete-modality and modality-missing testing on RGBNT201 and WMVEID863 (in %).

| RGBNT201 | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | Complete | | M (RGB) | | M (NIR) | | M (TIR) | | M (RGB+NIR) | | M (RGB+TIR) | | M (NIR+TIR) | | Average | |
| | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 |
| Baseline | 76.8 | 79.4 | 64.1 | 65.2 | 70.1 | 77.0 | 53.2 | 53.6 | 46.6 | 52.5 | 26.1 | 21.7 | 38.8 | 39.1 | 49.8 | 51.5 |
| +APD | 77.4 | 78.7 | 63.2 | 64.8 | 71.4 | 77.9 | 58.6 | 58.3 | 48.4 | 53.0 | 27.5 | 23.4 | 37.4 | 39.7 | 51.1 | 52.9 |
| +APD+HPI | **81.4** | **83.0** | **68.8** | **69.1** | **75.8** | **79.8** | **59.7** | **58.4** | **51.7** | **53.1** | **30.4** | **30.7** | **43.3** | **43.9** | **55.0** | **55.8** |
| WMVEID863 | | | | | | | | | | | | | | | | |
| Methods | Complete | | M (RGB) | | M (NIR) | | M (TIR) | | M (RGB+NIR) | | M (RGB+TIR) | | M (NIR+TIR) | | Average | |
| | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 |
| Baseline | 68.1 | 75.6 | 65.0 | 71.5 | 69.0 | 75.3 | 58.2 | 64.2 | 63.7 | 69.2 | 52.9 | 57.3 | 55.0 | 59.1 | 60.6 | 66.1 |
| +APD | 70.6 | **78.5** | 67.2 | 69.8 | **71.0** | 73.9 | 60.5 | 65.8 | **65.3** | 69.5 | 54.6 | 60.0 | 57.0 | 62.1 | 62.6 | 66.9 |
| +APD+HPI | **71.8** | 77.8 | **69.5** | **76.2** | 70.6 | **76.3** | **61.9** | **67.3** | 63.7 | **70.1** | **55.1** | **59.4** | **59.7** | **63.5** | **63.4** | **68.8** |

**Table 4**
Statistical stability analysis under complete-modality and fixed modality-missing settings. Results are reported as mean ± standard deviation over three random seeds.

| Method | RGBNT201 | | | | WMVEID863 | | | |
|---|---|---|---|---|---|---|---|---|
| | Complete | | Fixed Avg. | | Complete | | Fixed Avg. | |
| | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 |
| Baseline | 76.3 ± 3.3 | 78.7 ± 3.8 | 50.3 ± 2.3 | 49.7 ± 2.9 | 68.3 ± 0.6 | 76.2 ± 1.6 | 60.8 ± 0.2 | 66.3 ± 0.4 |
| +APD | 77.6 ± 1.5 | 79.9 ± 1.2 | 52.3 ± 1.0 | 52.3 ± 0.9 | 69.8 ± 0.7 | 76.6 ± 0.9 | 62.6 ± 0.4 | 67.8 ± 0.3 |
| +APD +HPI | 80.2 ± 1.7 | 82.0 ± 2.3 | 54.3 ± 1.0 | 54.6 ± 2.1 | 71.0 ± 0.4 | 77.0 ± 1.1 | 63.5 ± 0.5 | 68.5 ± 0.6 |

**Table 5**
Ablation and computational cost analysis on RGBNT201. Params, FLOPs, and latency are reported for each ablation variant. Fixed Avg. denotes the average performance over all fixed modality-missing settings.

| Modules | | Computational cost | | | RGBNT201 | | | |
|---|---|---|---|---|---|---|---|---|
| APD | HPI | Params (M) | FLOPs (G) | Latency (ms) | Complete | | Fixed Avg. | |
| | | | | | mAP | R-1 | mAP | R-1 |
| × | × | 86.5 | 22.1 | 34.0 | 76.8 | 79.4 | 49.8 | 51.5 |
| ✓ | × | 87.1 | 22.1 | 33.9 | 77.4 | 78.7 | 51.1 | 52.9 |
| ✓ | ✓ | 163.8 | 22.2 | 35.3 | **81.4** | **83.0** | **55.0** | **55.8** |

**Table 6**
Component analysis of APD and HPI on RGBNT201 and WMVEID863. Fixed Avg. denotes the average performance over all fixed modality-missing settings.

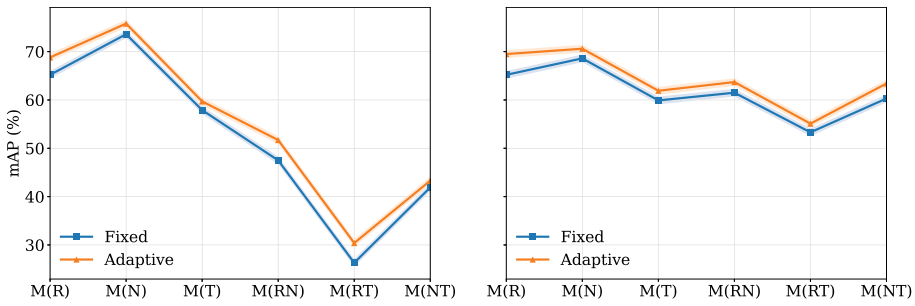
| APD | HPI | | RGBNT201 | | WMVEID863 | |
|---|---|---|---|---|---|---|
| CE | Stage I | Stage II | mAP | R-1 | mAP | R-1 |
| × | ✓ | ✓ | 53.5 | 54.2 | 61.8 | 67.8 |
| ✓ | ✓ | × | 54.0 | 54.4 | 61.2 | 66.0 |
| ✓ | × | ✓ | 49.9 | 49.1 | 62.2 | 68.2 |
| ✓ | ✓ | ✓ | **55.0** | **55.8** | **63.4** | **68.8** |

relatively marginal, which may be attributed to illumination-induced degradation of the available modalities and the resulting uncertainty in reconstructed features. Overall, APD and HPI improve the average modality-missing robustness and provide beneficial regularization for complete-modality representation learning.

**Stability Analysis of Key Modules.** To further evaluate the stability of the proposed modules, we repeat the experiments with three random seeds and report the mean ± standard deviation in Table 4. We report the results under the complete-modality setting and the fixed-missing average, where the fixed-missing average summarizes all six fixed modality-missing settings. As shown in Table 4, the APD variant and the complete PDRNet model generally achieve higher average performance than the baseline on both datasets. Although the results on RGBNT201 [3] show relatively larger variance, the overall improvement trend remains stable. In particular, the complete PDRNet model achieves the best fixed-missing average performance on both RGBNT201 and WMVEID863 [5], indicating that HPI provides stable benefits for improving modality-missing robustness.

**Efficiency Analysis of Ablation Variants.** To further evaluate the computational overhead of the proposed modules, Table 5 reports the Params, FLOPs, and inference latency of different ablation variants on RGBNT201 [3]. Compared with the baseline, APD only introduces a slight increase in parameters, while the FLOPs remain unchanged. Introducing HPI increases the parameter count due to its MLP- and FFN-based reconstruction modules. However, the increase in FLOPs is marginal, and the inference latency remains within a reasonable range. Moreover, the average performance under fixed modality-missing settings is clearly improved. These results indicate that HPI achieves a favorable balance between recognition performance and computational efficiency.

**Component Analysis of Key Modules.** To further analyze the contributions of the internal components in APD and HPI, we conduct component-level ablation experiments, as reported in Table 6. First, we evaluate the CE loss in APD, which constrains the identity discriminability of disentangled features and provides reliable identity-related content for reconstruction. Removing this loss leads to performance degradation, indicating its contribution to robust missing-modality reconstruction. We then analyze the two-stage design of HPI. Using only Stage I or Stage II performs worse than the complete HPI, showing




(a) RGBNT201 (b) WMVEID863

**Fig. 6.** Modality-missing experiments on the RGBNT201 and WMVEID863 using different prototype update strategies.

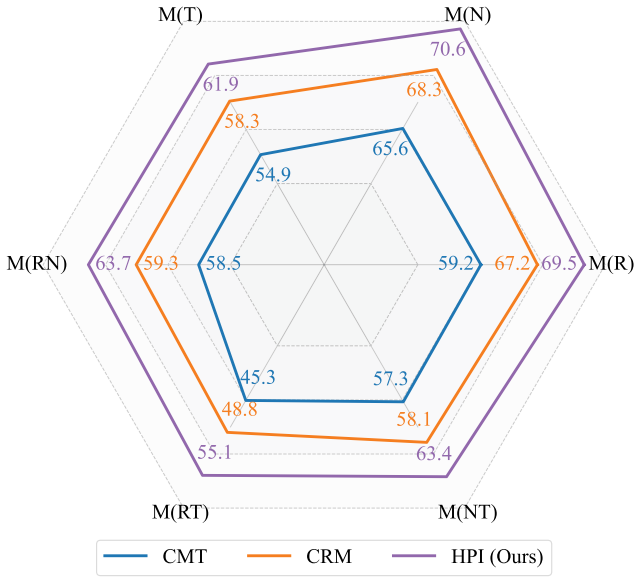



**Fig. 7.** Ablation experiments with different reconstruction methods on RGBNT201.

that the two stages are complementary: Stage I performs prototype-conditioned modulation, while Stage II further refines the reconstructed representation. Overall, the full design achieves the best fixed-missing average performance on both RGBNT201 [3] and WMVEID863 [5].

## 5. Discussions

### 5.1. Discussion on adaptive prototype update

To verify the effectiveness of the proposed adaptive prototype update in Adaptive Prototype-driven Disentanglement, we conduct a comparative study under two settings: fixed update weights and adaptive update weights. As shown in Fig. 6, the adaptive update consistently outperforms the fixed-weight update on both person and vehicle datasets. This indicates that, given the inherent differences in information content and reliability among samples within the same modality, the adaptive weighting strategy can better emphasize informative samples and suppress unreliable ones, thereby preserving discriminative cues and providing a more stable prototype basis for disentanglement.

### 5.2. Discussion on modality reconstruction

**Effects of Different Reconstruction Methods.** To validate the effectiveness of the proposed Hierarchical Prototype Injection (HPI) module, we compare it with representative reconstruction-based strategies. CMT [3] reconstructs missing modalities by directly translating the available-modality images into the missing-modality images at the input level, while CRM [12] performs reconstruction at the feature level by mapping available-modality features to missing-modality features. Although these methods operate on different forms of inputs, they all rely on direct transformation from the available modalities. As shown in Fig. 7, on the RGBNT201 [3] dataset, the proposed HPI produces robust reconstructed features via two-stage prototype injection and fusion, achieving the best performance compared with existing reconstruction methods.

**Effects of Prototype Number.** In this section, we analyze the effect of modality-specific prototypes on missing-modality reconstruction. We compare the reconstruction without prototype guidance ($K = 0$) and the reconstruction with different numbers of prototypes, where $K$ denotes the number of prototypes per modality. As shown in Table 7, introducing prototypes improves the average performance over $K = 0$, indicating that modality-level prototype guidance is beneficial for reconstruction. Among different settings, $K = 1$ achieves the best average performance on both RGBNT201 [3] and WMVEID863 [5]. Although multiple prototypes can describe more fine-grained modality variations, they may also absorb identity-discriminative details into modality-specific components and weaken the residual identity representation. Therefore, one prototype per modality provides a more stable modality-level prior and better balances modality disentanglement and identity preservation. It should be clarified that the choice of one prototype per modality is not selected on the validation set for each dataset or missing-modality condition. Instead, based on the sensitivity analysis, we empirically fix $K = 1$ as the default setting and use it throughout all experiments.

**Effects of Different Loss Functions.** Fig. 8 compares the mAP performance of reconstruction losses under modality-missing settings, including the reconstruction loss ($\mathcal{L}_{cr}$) [12], the cosine similarity loss ($\mathcal{L}_{cos}$), and our reconstruction consistency loss ($\mathcal{L}_{RC}$). Here, $\mathcal{L}_{cr}$ is an MSE-based reconstruction constraint. The results show that $\mathcal{L}_{RC}$ consistently achieves the best performance on both RGBNT201 [3] and WMVEID863 [5]. Compared with $\mathcal{L}_{cr}$, which focuses on numerical proximity, and $\mathcal{L}_{cos}$, which focuses on directional similarity, $\mathcal{L}_{RC}$ jointly enforces distance-level consistency and directional similarity. This joint constraint better preserves identity-discriminative cues and reduces semantic drift caused by pseudo-reconstruction, leading to more robust missing-modality representation recovery.

### 5.3. Discussion on hyperparameters

Our method mainly involves two loss weights in Eq. (20): $\lambda_1$ and $\lambda_2$. $\lambda_1$ weights the disentanglement loss, while $\lambda_2$ controls the alignment strength between reconstructed and real features. Too large $\lambda_1$ may harm identity discriminability, and too large $\lambda_2$ may over-align modalities and degrade performance. We vary both from 0.1 to 1.0 with a step of 0.2, and Table 8 reports the average mAP and Rank-1 over all fixed modality-missing settings. The best overall performance

**Table 7**
Ablation results of the modality-specific prototype number on WMVEID863 and RGBNT201 datasets (in %). $K$ denotes the number of prototypes per modality.

| Missing modality | WMVEID863 | | | | | RGBNT201 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $K=0$ | $K=1$ | $K=2$ | $K=4$ | $K=6$ | $K=0$ | $K=1$ | $K=2$ | $K=4$ | $K=6$ |
| M (RGB) | 68.1 | 69.5 | 67.8 | 68.2 | **69.6** | 67.5 | 68.8 | 67.5 | **68.9** | 67.1 |
| M (NIR) | 69.5 | **70.6** | 69.3 | 69.9 | 69.9 | 74.7 | **75.8** | 73.6 | 74.5 | 72.5 |
| M (TIR) | 61.5 | 61.9 | **62.2** | 61.7 | 61.6 | 57.7 | **59.7** | 56.2 | 56.8 | 56.3 |
| M (RGB+NIR) | 61.6 | **63.7** | 61.2 | 61.6 | 61.2 | **52.0** | 51.7 | 48.2 | 51.1 | 47.7 |
| M (RGB+TIR) | 53.8 | **55.1** | 54.1 | 52.9 | 54.7 | 28.6 | **30.4** | 26.8 | 28.1 | 27.4 |
| M (NIR+TIR) | **60.1** | 59.7 | 59.4 | 57.8 | 57.1 | 42.7 | **43.3** | 40.8 | 42.2 | 41.6 |
| Average | 62.4 | **63.4** | 62.3 | 62.0 | 62.4 | 53.9 | **55.0** | 52.2 | 53.6 | 52.1 |

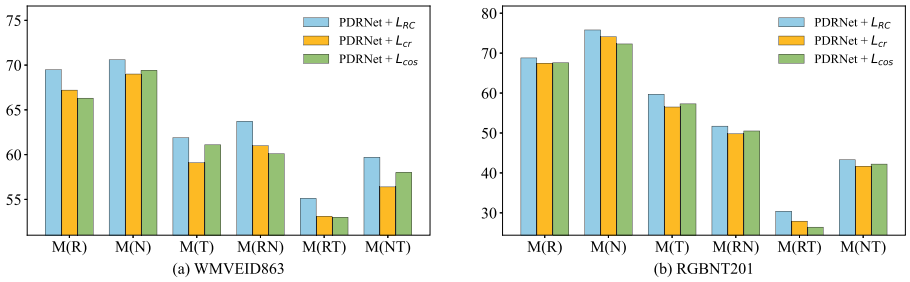



**Fig. 8.** Ablation study comparing the proposed $\mathcal{L}_{RC}$, $\mathcal{L}_{cr}$ and $\mathcal{L}_{cos}$ on RGBNT201 and WMVEID863.

**Table 8**
Hyperparameter analysis on RGBNT201 and WMVEID863.

| Hyper-parameter | Value | RGBNT201 | | WMVEID863 | |
|---|---|---|---|---|---|
| | | mAP | R-1 | mAP | R-1 |
| $\alpha$ | 0.9 | 51.9 | 52.5 | 61.6 | 66.5 |
| | 0.99 | **55.0** | **55.8** | **63.4** | **68.8** |
| | 0.999 | 53.2 | 55.7 | 62.5 | 67.2 |
| $\lambda_1$ ($\lambda_2 = 0.1$) | 0.1 | **55.0** | **55.8** | **63.4** | 68.8 |
| | 0.3 | 54.6 | 57.7 | 63.2 | **69.1** |
| | 0.5 | 52.3 | 55.2 | 62.7 | 66.3 |
| | 0.7 | 50.4 | 52.7 | 60.5 | 63.5 |
| | 0.9 | 48.7 | 49.6 | 61.8 | 65.9 |
| | 1.0 | 45.2 | 46.1 | 59.7 | 60.3 |
| $\lambda_2$ ($\lambda_1 = 0.1$) | 0.1 | 55.0 | **55.8** | 63.4 | **68.8** |
| | 0.3 | **55.3** | 57.7 | 62.7 | 68.5 |
| | 0.5 | 54.3 | 56.2 | 63.1 | 67.4 |
| | 0.7 | 54.4 | 55.8 | **64.5** | 67.8 |
| | 0.9 | 53.8 | 55.9 | 62.8 | 67.6 |
| | 1.0 | 52.2 | 54.6 | 60.5 | 63.9 |

is achieved at $\lambda_1 = \lambda_2 = 0.1$, which we use in all experiments. We also analyze the fixed update weight $\alpha$ in the prototype warm-up stage. Since the prototypes are initialized as zero vectors, $\alpha$ is used in the first epoch to bootstrap prototype representations. As shown in Table 8, $\alpha = 0.99$ achieves the best overall performance on both RGBNT201 [3] and WMVEID863 [5], indicating that a moderate warm-up update strength is more suitable for stable prototype initialization. Therefore, we adopt $\alpha = 0.99$ as the default setting.

### 5.4. Discussion on visualization

**Feature Distributions.** To validate the effectiveness of PDRNet, we visualize the t-SNE [41] projections of features produced by the baseline model and by the model equipped with our proposed modules. As shown in Fig. 9(a), (b), and (d), the original identity features from the baseline exhibit a clear modality gap, indicating pronounced cross-modality distribution shift. After incorporating our modules, the feature clusters become noticeably more compact and the inter-modality gap is substantially reduced, demonstrating that PDRNet effectively enhances multi-modal feature extraction and promotes cross-modality alignment. Moreover, we visualize the disentangled identity features in Fig. 9(c). Compared with the original features in Fig. 9(b), the disentangled multi-modal features are more tightly clustered, suggesting that our method can reduce modality-specific cues while preserving identity-related discriminative information.

**Disentanglement Visualization.** To further validate the disentanglement effect of APD, we visualize the original features and residual features with modality labels using t-SNE on RGBNT201 [3] and WMVEID863 [5]. As shown in Fig. 10, (a) and (b) show the results on RGBNT201, while (c) and (d) show the results on WMVEID863. Specifically, (a) and (c) denote the original features, and (b) and (d) denote the disentangled residual features. On RGBNT201, the TIR modality in the original feature space shows relatively clear modality-wise separation from the other modalities, whereas the residual features exhibit weaker modality separation and stronger cross-modality mixing. This phenomenon is more obvious on WMVEID863, where the original features present clear cross-modality separation, while the residual features become more mixed across modalities. These observations provide qualitative evidence that APD can reduce modality-specific cues in the residual features and preserve more modality-invariant identity-discriminative information.

**Area of Concern Visualization.** As shown in Fig. 11, we visualize the Grad-CAM [17] responses of PDRNet on RGBNT201 [3] and WMVEID863 [5]. Fig. 11(a) shows the inputs, and Fig. 11(b) presents attention maps under the complete-modality setting, where responses consistently focus on identity-discriminative vehicle regions across modalities. Fig. 11(c)–(e) report cases with RGB/NIR/TIR missing. By reconstructing the missing modality using the remaining two modalities and the corresponding prototype, PDRNet still highlights informative regions, indicating effective compensation for missing information. Fig. 11(f)–(h) further demonstrate that even in extreme missing-modality scenarios, PDRNet can preserve part of the discriminative regions, showing strong robustness. Fig. 11(i)–(p) provide analogous results on the person Re-ID dataset, where PDRNet maintains meaningful attention under all missing-modality conditions. Overall, these visualizations qualitatively validate PDRNet's effectiveness and generalization for modality-missing object Re-ID.

## 6. Conclusion

In this work, we propose a modality-specific prototypes for disentangled reconstruction framework for object Re-ID. The framework

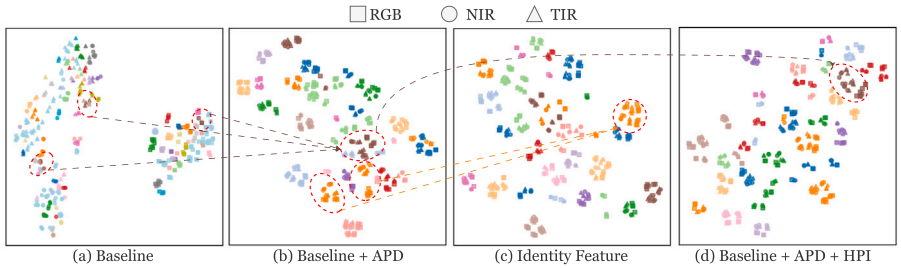



**Fig. 9.** Visualization of t-SNE on RGBNT201. (a) Baseline Feature Distributions. (b) Baseline + APD Feature Distributions. (c) Distributions of Residual Identity Features. (d) Baseline + APD + HPI Feature Distributions.

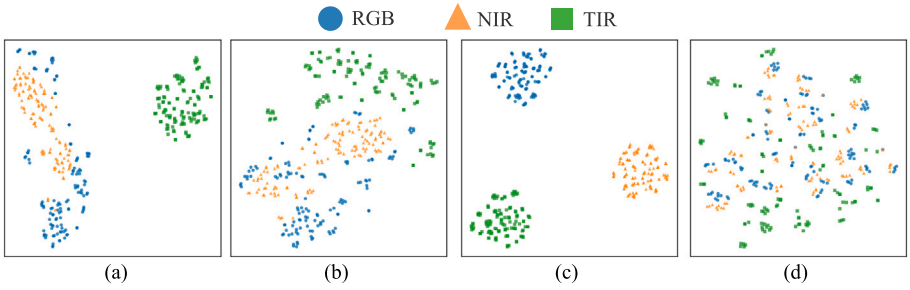



**Fig. 10.** Disentanglement visualization on RGBNT201 and WMVEID863. (a) and (b) show the results on RGBNT201, while (c) and (d) show the results on WMVEID863. (a) and (c) denote the original features, and (b) and (d) denote the disentangled residual features.

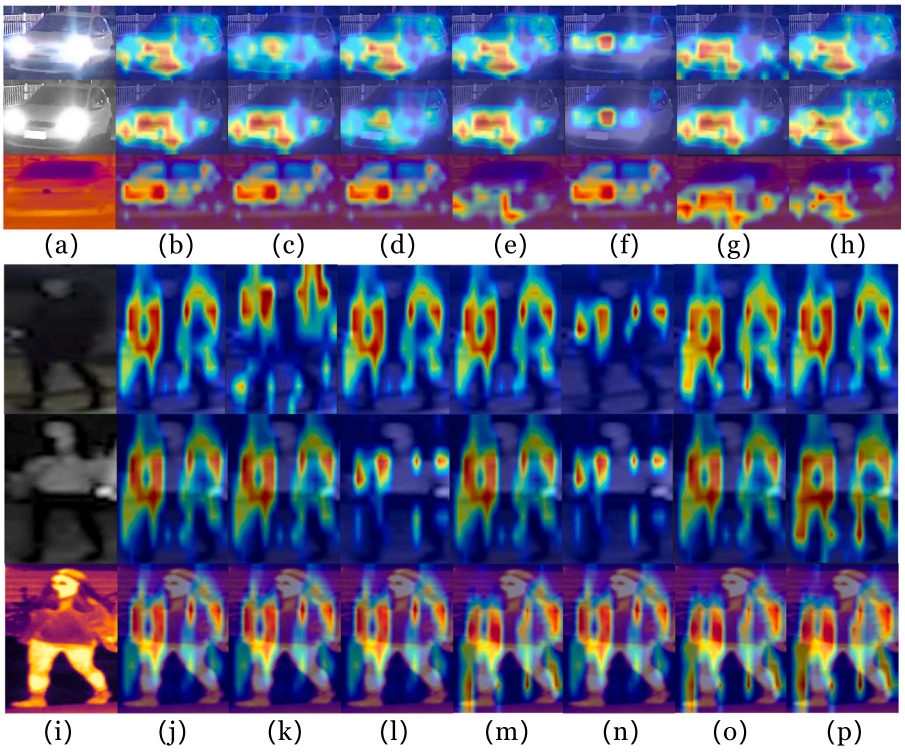



**Fig. 11.** The visualized results of the attention map. (a) Inputs, (b) complete, (c) M (RGB), (d) M (NIR), (e) M (TIR), (f) M (RGB+NIR), (g) M (RGB+TIR), (h) M (NIR+TIR). (i)–(p) show the corresponding results on the RGBNT201 dataset under the same missing-modality settings.

addresses the modality-missing problem, where entanglement between modality-specific information and identity information leads to biased and inconsistent representations. We develop an Adaptive Prototype-driven Disentanglement module to encourage the disentanglement between discriminative identity cues and generic modality information while adaptively updating modality prototypes, and a Hierarchical Prototype Injection module to enable controllable and robust missing-modality reconstruction through two-stage prototype injection and progressive fusion. Extensive experiments on four multi-modal datasets demonstrate the effectiveness and robustness of our method under both complete-modality and modality-missing settings. Despite leveraging modality-specific prototypes for feature completion, discriminative

cues that are unique to the missing modality are not yet fully exploited. In future work, we will explore reconstruction strategies that better preserve the missing modality's distinctive discriminative characteristics in the reconstructed representations.

## CRediT authorship contribution statement

**Zhendong Xu:** Writing – original draft, Investigation. **Aihua Zheng:** Methodology, Conceptualization. **Zi Wang:** Writing – review & editing. **Chenglong Li:** Formal analysis, Data curation. **Dong Geng:** Validation, Data curation. **Jin Tang:** Resources, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Aihua Zheng reports was provided by Anhui University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This research is supported in part by the National Natural Science Foundation of China under Grant 62372003, the Natural Science Foundation of Anhui Province under Grant 2308085Y40. The authors acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

## Data availability

Data will be made available on request.

## References

[1] Y. Xie, H. Wu, Y. Lin, J. Zhu, H. Zeng, Pairwise difference relational distillation for object re-identification, Pattern Recognit. 152 (2024) 110455.
[2] D. Han, B. Liu, S. Shao, W. Liu, Y. Zhou, Feature aggregation and connectivity for object re-identification, Pattern Recognit. 157 (2025) 110869.
[3] A. Zheng, Z. Wang, Z. Chen, C. Li, J. Tang, Robust multi-modality person re-identification, in: AAAI, 2021, pp. 3529–3537.
[4] A. Zheng, X. Zhu, Z. Ma, C. Li, J. Tang, J. Ma, Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark, Inf. Fusion 100 (2023) 101901.
[5] A. Zheng, Z. Ma, Y. Sun, Z. Wang, C. Li, J. Tang, Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification, Inf. Fusion 116 (2025) 102800.
[6] X. Zheng, X. Huang, C. Ji, X. Yang, P. Sha, L. Cheng, Multi-modal person re-identification based on transformer relational regularization, Inf. Fusion 103 (2024) 102128.
[7] Y. Wang, Y. Lv, P. Zhang, H. Lu, Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification, in: CVPR, 2025, pp. 29701–29710.
[8] Y. Feng, F. Chen, G. Sun, F. Wu, Y. Ji, T. Liu, S. Liu, X.-Y. Jing, J. Luo, Learning multi-granularity representation with transformer for visible-infrared person re-identification, Pattern Recognit. 164 (2025) 111510.
[9] Y. Feng, F. Chen, J. Yu, Y. Ji, F. Wu, S. Liu, X.-Y. Jing, Homogeneous and heterogeneous relational graph for visible-infrared person re-identification, Pattern Recognit. 158 (2025) 110981.
[10] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, G. Carneiro, Multi-modal learning with missing modality via shared-specific feature modelling, in: CVPR, 2023, pp. 15878–15887.
[11] Y. Zhang, F. Liu, X. Zhuang, Y. Hou, Y. Zhang, Prototype-based sample-weighted distillation unified framework adapted to missing modality sentiment analysis, Neural Netw. 177 (2024) 106397.
[12] Y. Wang, X. Liu, P. Zhang, H. Lu, Z. Tu, H. Lu, TOP-ReID: Multi-spectral object re-identification with token permutation, in: AAAI, 2024, pp. 5758–5766.
[13] S. Zhang, W. Luo, D. Cheng, Y. Xing, G. Liang, P. Wang, Y. Zhang, Prompt-based modality alignment for effective multi-modal object re-identification, IEEE Trans. Image Process. 34 (2025) 2450–2462.
[14] Y. Dai, H. Chen, J. Du, R. Wang, S. Chen, H. Wang, C.-H. Lee, A study of dropout-induced modality bias on robustness to missing video frames for audio-visual speech recognition, in: CVPR, 2024, pp. 27445–27455.
[15] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, L. He, Multi-level feature learning for contrastive multi-view clustering, in: CVPR, 2022, pp. 16051–16060.
[16] J. Feng, A. Wu, W.-S. Zheng, Shape-erased feature learning for visible-infrared person re-identification, in: CVPR, 2023, pp. 22752–22761.
[17] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: ICCV, 2017, pp. 618–626.
[18] X. Hao, S. Zhao, M. Ye, J. Shen, Cross-modality person re-identification via modality confusion and center aggregation, in: ICCV, 2021, pp. 16403–16412.
[19] Z. Wei, X. Yang, N. Wang, X. Gao, Dual-adversarial representation disentanglement for visible infrared person re-identification, IEEE Trans. Inf. Forensics Secur. 19 (2023) 2186–2200.
[20] H. Li, C. Li, X. Zhu, A. Zheng, B. Luo, Multi-spectral vehicle re-identification: A challenge, in: AAAI, 2020, pp. 11345–11353.
[21] Z. Wang, C. Li, A. Zheng, R. He, J. Tang, Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification, in: AAAI, 2022, pp. 2633–2641.
[22] Z. Wang, H. Huang, A. Zheng, R. He, Heterogeneous test-time training for multi-modal person re-identification, in: AAAI, 2024, pp. 5850–5858.
[23] P. Zhang, Y. Wang, Y. Liu, Z. Tu, H. Lu, Magic tokens: Select diverse tokens for multi-modal object re-identification, in: CVPR, 2024, pp. 17117–17126.
[24] Z. Yu, Z. Huang, M. Hou, J. Pei, Y. Yan, Y. Liu, D. Sun, Representation selective coupling via token sparsification for multi-spectral object re-identification, IEEE T. Circuits Syst. Video Technol. 35 (2024) 3633–3648.
[25] Y. Wang, X. Liu, T. Yan, Y. Liu, A. Zheng, P. Zhang, H. Lu, Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt, in: AAAI, 2025, pp. 8150–8158.
[26] S. Li, C. Li, A. Zheng, J. Tang, B. Luo, ICPL-ReID: Identity-conditional prompt learning for multi-spectral object re-identification, IEEE Trans. Multimed. 28 (2025) 599–613.
[27] T. Zhou, Feature fusion and latent feature learning guided brain tumor segmentation and missing modality recovery network, Pattern Recognit. 141 (2023) 109665.
[28] Z. Lian, L. Chen, L. Sun, B. Liu, J. Tao, GCNet: Graph completion network for incomplete multimodal learning in conversation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 8419–8432.
[29] H. Liu, J. Gu, Z. Li, M. Wang, Q.J. Wu, W. Jiang, CoMix: Collaborative mixed learning via style fuzzy normalization for visible–Infrared person re-identification, IEEE Trans. Syst. Man Cybern.: Syst. (2025).
[30] J. Yao, J. Zhang, Y. Wang, L. Zhuo, Asymmetric simulation-enhanced flow reconstruction for incomplete multimodal learning, Pattern Recognit. (2025) 112413.
[31] Y. Wang, Y. Liu, A. Zheng, P. Zhang, Decoupled feature-based mixture of experts for multi-modal object re-identification, in: AAAI, 2025, pp. 8141–8149.
[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv:2010.11929.
[33] A. Bardes, J. Ponce, Y. LeCun, Vicreg: Variance-invariance-covariance regularization for self-supervised learning, in: ICLR, 2022.
[34] Y. Shigeto, M. Shimbo, Y. Yoshikawa, A. Takeuchi, Learning decorrelated representations efficiently using fast fourier transform, in: CVPR, 2023, pp. 2052–2060.
[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: CVPR, 2016, pp. 2818–2826.
[36] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, 2017, arXiv:1703.07737.
[37] W. Dong, X. Yang, D. Cheng, N. Wang, X. Gao, Escaping modal interactions: An efficient DESANet for multi-modal object re-identification, IEEE Trans. Image Process. 34 (2025) 5068–5083.
[38] Y. Feng, J. Li, C. Xie, L. Tan, J. Ji, Multi-modal object re-identification via sparse mixture-of-experts, in: ICML, 2025.
[39] Y. Feng, J. Li, J. Hu, Y. Zhang, L. Tan, J. Ji, MDReID: Modality-decoupled learning for any-to-any multi-modal object re-identification, in: NeurIPS, 2025.
[40] R. Xi, Miss-ReID: Delivering robust multi-modality object re-identification despite missing modalities, in: NeurIPS, 2025.
[41] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.