CMCC-ReID: Cross-Modality Clothing-Change Person Re-Identification

Haoxuan Xu[1], Hanzi Wang[2], and Guanglin Niu[1⋆]

[1] School of Artificial Intelligence, Beihang University, Beijing, China
[2] School of Informatics, Xiamen University, Xiamen, Fujian, China
{xhaoxuan,beihangngl}@buaa.edu.cn
https://github.com/Xuan266/CMCC-ReID

Abstract. Person Re-Identification (ReID) faces severe challenges from modality discrepancy and clothing variation in long-term surveillance scenario. While existing studies have made significant progress in either Visible-Infrared ReID (VI-ReID) or Clothing-Change ReID (CC-ReID), real-world surveillance system often face both challenges simultaneously. To address this overlooked yet realistic problem, we define a new task, termed Cross-Modality Clothing-Change Re-Identification (CMCC-ReID), which targets pedestrian matching across variations in both modality and clothing. To advance research in this direction, we construct a new benchmark SYSU-CMCC, where each identity is captured in both visible and infrared domains with distinct outfits, reflecting the dual heterogeneity of long-term surveillance. To tackle CMCC-ReID, we propose a Progressive Identity Alignment Network (PIA) that progressively mitigates the issues of clothing variation and modality discrepancy. Specifically, a Dual-Branch Disentangling Learning (DBDL) module separates identity-related cues from clothing-related factors to achieve clothing-agnostic representation, and a Bi-Directional Prototype Learning (BPL) module performs intra-modality and inter-modality contrast in the embedding space to bridge the modality gap while further suppressing clothing interference. Extensive experiments on the SYSU-CMCC dataset demonstrate that PIA establishes a strong baseline for this new task and significantly outperforms existing methods.

Keywords: Person re-identification · Cross-modality · Clothing-change

1 Introduction

Person Re-Identification (ReID) [9, 34, 59, 63, 69] aims to match pedestrian images captured across different non-overlapping cameras, and has become a fundamental task in intelligent surveillance and public security applications. Despite the remarkable progress achieved on controlled benchmarks, traditional ReID methods struggle to generalize well in unconstrained real-world environments, particularly under long-term surveillance conditions.

⋆ Corresponding author.
This work is partially supported by the National Natural Science Foundation of China (No. U25A20531 and No. 62376016).

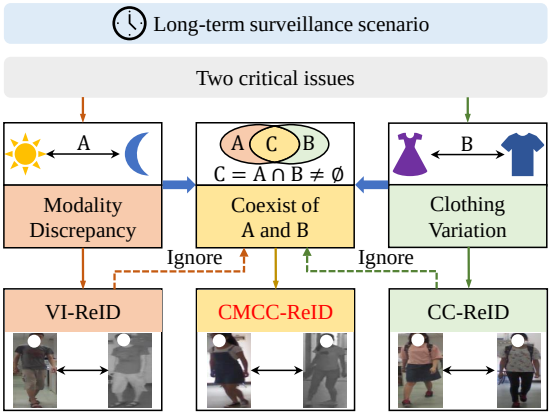


Fig. 1: Motivation of CMCC-ReID. In the long-term surveillance scenarios, two critical issues (*i.e.*, modality discrepancy and clothing variation) are not mutually exclusive but coexist. This observation motivates the CMCC-ReID task, which targets pedestrian matching across both modality and clothing changes.

In such long-term surveillance scenario, two critical issues jointly challenge conventional ReID systems: active clothing variations resulting from intentional outfit changes and passive modality discrepancies induced by illumination transitions between day and night. As illustrated in Fig. 1, existing research typically isolates these challenges into two separate settings: Visible–Infrared Re-ID (VI-ReID) [47, 66] assumes fixed clothing conditions, while Clothing-Change Re-ID (CC-ReID) [37, 55] assumes consistent good illumination. However, these assumptions are overly idealized, as they overlook the fact that the two issues are not mutually exclusive and often coexist in real-world scenarios. This observation motivates us to investigate a more realistic and challenging setting, termed **C**ross-**M**odality **C**lothing-**C**hange **Re-ID** (**CMCC-ReID**), which requires jointly addressing both modality discrepancy and clothing variation for reliable long-term identity association.

To advance research on the CMCC-ReID task, we construct a new dataset, termed **SYSU C**ross-**M**odality **C**lothing-**C**hange (**SYSU-CMCC**), derived from PRCC [55] and SYSU-MM01 [47]. In SYSU-CMCC, each identity is captured under both visible and infrared modalities, ensuring the presence of modality discrepancies. Moreover, to introduce clothing variations, the visible and infrared images of the same identity are collected under different outfits. The detailed description of the SYSU-CMCC dataset is provided in Sec. 3.

The core challenge of CMCC-ReID lies in learning identity-consistent representations under the coexistence of modality discrepancy and clothing variation. Unlike a simple combination of CC-ReID and VI-ReID, CMCC-ReID presents a more entangled setting where these two types of heterogeneity mutually reinforce each other. Beyond the inherent clothing variation and modality gap inherited

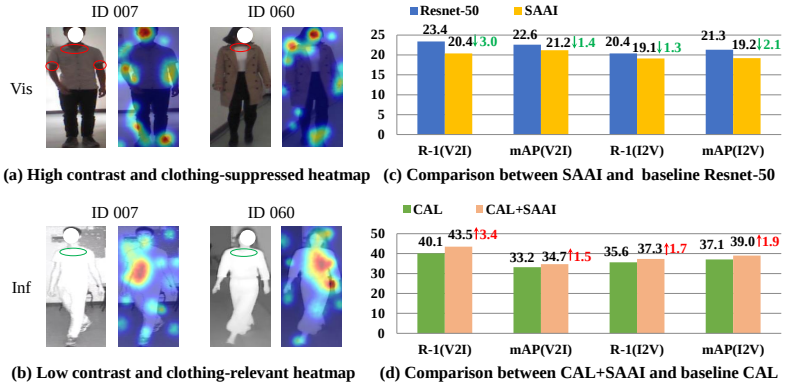


Fig. 2: Illustration of the key issues in CMCC-ReID. (a) In the visible modality, the high contrast enables CAL to produce a heatmap with suppressed clothing cues. (b) In the infrared modality, low contrast causes CAL to generate a heatmap dominated by clothing information. (c) Comparison between SAAI and its ResNet-50 baseline shows that direct modality alignment may even degrade performance under clothing variation. (d) Integrating SAAI into CAL yields moderate improvements once clothing interference is mitigated.

from the two tasks, CMCC-ReID further introduces two key issues that must be addressed:

(1) Extracting clothing-irrelevant features under information-degraded infrared modality. Existing CC-ReID methods [8,14,26] typically learn clothing-invariant representations under the visible modality with rich color and texture cues. However, in the infrared modality, the absence of such cues causes the body to appear uniformly bright, resulting in low visual contrast between clothing regions and other body parts. This degradation blurs semantic boundaries and hampers the disentanglement of identity cues from clothing information. To verify this, we visualize the attention maps of CAL [8] under both modalities. As shown in Fig. 2(a), CAL effectively suppresses clothing-related responses when high visual contrast is present. In contrast, under the infrared modality, as shown in Fig. 2(b), its attention shifts toward clothing-related regions, hindering the extraction of identity-consistent features. Consequently, although existing methods perform well under the visible spectrum, they fail to generalize to the information-degraded infrared modality, leading to limited performance.

(2) Preventing clothing interference before and during cross-modality alignment. Existing VI-ReID methods typically eliminate modality discrepancy by aligning global [45,67] or part-level [6,49] representations within shared embedding space. However, in the CMCC-ReID, clothing acts as a noisy factor, performing direct feature alignment without suppressing clothing-related interference can mislead the model to align clothing-related clues instead of true identity semantics. To empirically validate this, we compare a state-of-the-art VI-ReID method SAAI [6] with its baseline (*i.e.*, a standard ResNet-50 [12] network

trained solely with an identity classification loss). As shown in Fig. 2(c), SAAI performs even inferior to the baseline, suggesting that naive modality alignment may be counterproductive. Whereas, this does not imply that modality alignment is ineffective. When the core module of SAAI is incorporated into CAL [8], the performance exhibits a moderate improvement as shown in Fig. 2(d). These observations indicate that while modality alignment alone may strengthen appearance bias, it becomes beneficial when conducted upon clothing-irrelevant representations. Therefore, suppressing clothing interference before and during cross-modality alignment is crucial for CMCC-ReID.

To this end, we propose a novel **P**rogressive **I**dentity **A**lignment Network (**PIA**) that progressively disentangles and aligns identity representations across both clothing and modality variations. The central idea of PIA lies in a progressive learning paradigm that first constructs clothing-irrelevant identity representations in Stage I and subsequently aligns them across modalities in Stage II. In Stage I, a **D**ual-**B**ranch **D**isentanglement **L**earning (**DBDL**) module is proposed to explicitly separate identity-related cues from clothing-related noisy through a dual-branch architecture combined with an orthogonality constraint. In Stage II, a **B**i-Directional **P**rototype **L**earning (**BPL**) module is introduced to jointly perform intra-modality contrast to further refine identity consistency and inter-modality contrast to reduce the modality gap within a prototype learning scheme. By jointly optimizing the DBDL and BPL modules under the progressive learning strategy, PIA evolves from disentangled feature extraction to modality alignment, achieving consistent and discriminative identity representations.

Our main contributions are summarized as follows:

- We introduce a new and realistic task CMCC-ReID, which jointly addresses clothing variation and modality discrepancy. To advance this study, we construct the SYSU-CMCC dataset, where each identity is captured under distinct modalities and disjoint clothing conditions.
- We propose PIA for effective CMCC-ReID, focusing on extracting identity-consistent representations through a progressive learning paradigm.
- We propose a DBDL module for clothing-irrelevant feature extraction in both modalities and a BPL module for effective cross-modality alignment while further suppressing clothing interference.
- Extensive experiments validate that PIA establishes a strong baseline and significantly outperforms existing CC-ReID and VI-ReID methods.

2 Related Work

2.1 Visible-Infrared Person ReID

VI-ReID [47, 52, 66] aims to match pedestrian images across different modalities. To mitigate the modality gap, extensive studies [7, 38, 39, 50, 54, 57, 58] have explored learning modality-shared representations through various strategies, including auxiliary modality construction [23, 25, 32, 45, 67], feature alignment [2, 18, 21, 42, 48], modality compensation [16, 33, 60, 62, 64, 65], metric learn-

ing [3, 11, 31], and generative translation [4, 41, 43, 68]. Moreover, several part-based approaches [6, 48, 49] further enhance cross-modality discrimination by enforcing local semantic consistency, thereby reducing intra-class variance. However, existing VI-ReID methods are developed under the assumption of clothing consistency, overlooking the fact that clothing variation and modality discrepancy often coexist in real-world scenarios, which limits their practical applicability. Moreover, conventional VI-ReID approaches that focus primarily on modality alignment cannot be directly extended to the CMCC-ReID task. Without explicit constraints to suppress clothing interference, these methods tend to overfit to appearance-specific patterns, leading to unstable identity representations across modalities and clothing conditions.

A recent study [44] extends VI-ReID to clothing-change scenarios but treats clothing variation as a supplementary factor rather than systematically analyzing its coupling with modality discrepancy. Furthermore, their method relies on lightweight adaptations of existing VI-ReID baselines, whereas our PIA introduces a progressive learning framework incorporating DBDL and BPL to disentangle modality and clothing interference in a stage-wise manner.

2.2 Clothing-Change Person ReID

CC-ReID [37,40,51,53,55,61] aims to retrieve pedestrian images across different outfits. To extract clothing-irrelevant representations, existing methods [10, 13, 26–28, 36] enhance identity discrimination through either multi-modality guidance or single-modality disentanglement. The former exploits auxiliary cues such as sketches [55], keypoints [37], human contours [14], gait patterns [19], or 3D body geometry [1, 29, 30] to mitigate clothing interference. The latter focuses on learning clothing-agnostic features within a single modality via adversarial learning [8], status awareness [15], or causal inference [56]. Beyond the limited applicability similar to VI-ReID, existing CC-ReID methods cannot be directly generalized to the CMCC-ReID task. Due to the information degradation in the infrared modality, these methods struggle to suppress clothing interference, resulting in suboptimal performance under cross-modality and clothing-change conditions.

3 The SYSU-CMCC Dataset

To advance research on the CMCC-ReID task, we construct a new dataset named SYSU-CMCC, which combines the heterogeneous characteristics of visible and infrared modalities with clothing variations within the same identity.

Dataset Construction. The SYSU-CMCC dataset is constructed by integrating PRCC and SYSU-MM01, where visible images from PRCC and infrared images from SYSU-MM01 represent two distinct clothing appearances per identity. Since both source datasets inevitably contain **annotation noise** and **low-quality samples**, and share overlapping identity labels, we adopt a manual cross-dataset verification protocol to ensure identity consistency. As illustrated

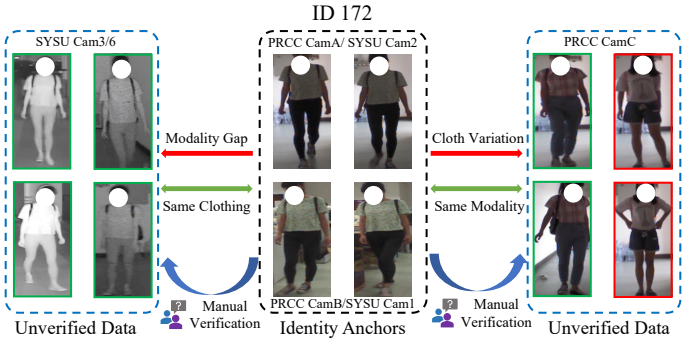


Fig. 3: Cross-dataset verification protocol for SYSU-CMCC construction. Samples from PRCC camA/B and SYSU-MM01 cam1/2 serve as **identity anchors** due to their identical content. For remaining samples (PRCC camC and SYSU-MM01 cam3/6), manual verification is performed: inconsistent samples are marked with red bounding boxes and discarded, while consistent samples are highlighted with green bounding boxes and retained to form the SYSU-CMCC dataset.

Table 1: Statistical comparisons between the SYSU-CMCC dataset with existing CC-ReID and VI-ReID datasets.

Dataset	Clothing-change	Cross-modality	IDs	Images	Cameras
PRCC [55]	✔	✗	221	33,698	3
LTCC [37]	✔	✗	152	17,138	12
RegDB [35]	✗	✔	412	8240	2
SYSU-MM01 [47]	✗	✔	491	38,271	6
SYSU-CMCC	✔	✔	214	18,375	3

in Fig. 3, samples from PRCC camA/B and SYSU-MM01 cam1/2 are identical, serving as reliable identity anchors across the two datasets. Building upon these anchors, we further perform manual verification on samples from PRCC camC and SYSU-MM01 cam3/6. Samples violating identity consistency are marked with red bounding boxes and removed, while those satisfying consistency are highlighted with green bounding boxes and retained to construct the final SYSU-CMCC dataset. This construction ensures that each identity simultaneously exhibits modality discrepancy and clothing variation, while the rigorous filtering process significantly improves annotation quality and dataset reliability.

Dataset Description and Evaluation Protocol. As summarized in Tab. 1, SYSU-CMCC comprises 214 identities and 18,375 images, encompassing diverse clothing styles, illumination conditions, and camera viewpoints. Specifically, pedestrians captured by Cam1/2 are under the infrared modality wearing clothing A, while the same identities in Cam3 are imaged under the visible modality with clothing B. During testing, we adopt the Visible to Infrared (V2I) and Infrared to Visible (I2V) matching protocols to evaluate cross-modality retrieval under clothing variation. The SYSU-CMCC dataset provides the first

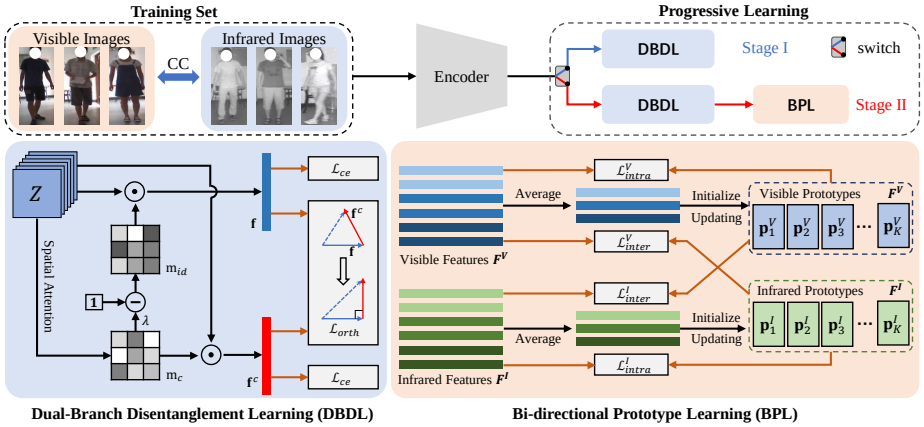


Fig. 4: Overview of PIA, which consists of two key components: a Dual-Branch Disentanglement Learning (DBDL) module for clothing-invariant feature extraction in both modalities, and a Bi-Directional Prototype Learning (BPL) module for cross-modality identity alignment. Through progressive optimization, PIA achieves robust and modality-consistent identity representation.

benchmark for jointly studying modality discrepancy and clothing variation, offering a platform for developing CMCC-ReID.

4 Methodology

In this section, we present the details of our proposed PIA. As illustrated in Fig. 4, PIA is built upon a convolutional backbone. To effectively suppress clothing interference before cross-modality alignment, which is crucial for CMCC-ReID, PIA adopts a progressive learning paradigm that first mitigates clothing variation and then bridges modality discrepancies. Moreover, PIA comprises two key components: a DBDL module that extracts identity-consistent and clothing-agnostic representations, participating in both Stages, and a BPL module that aligns identity embeddings across modalities, which is introduced in Stage II.

4.1 Dual-Branch Disentanglement Learning

In CMCC-ReID, both visible and infrared modalities inevitably contain clothing-related cues that interfere with identity recognition. Regardless of the modality, identity and clothing can be regarded as two semantically independent factors: the appearance of clothing should not influence the inherent identity representation, and changes in identity should not affect the representation of the same outfit. Based on this insight, we propose the DBDL module, which explicitly models identity and clothing features through dual-branch feature learning. Furthermore, it enforces orthogonality between the identity and clothing feature

subspaces, ensuring that identity features and clothing features do not mutually interfere with each other.

Formally, given a training sample x_i, we denote its identity label as y_i^{id} and its clothing label as y_i^c, where the latter one is defined as a fine-grained class following [8]. For an input image x_i, its global feature map $\mathbf{Z} \in \mathbb{R}^{C\times H\times W}$ extracted by the shared convolutional backbone serves as the input to DBDL. To guide the clothing branch in effectively capturing clothing-related cues, we incorporate a lightweight spatial attention mechanism [46] to generate a clothing-related spatial weight mask $\mathbf{m}_c$, formulated as:

$$\mathbf{m}_c = \sigma\left(\mathbf{W}_1 * [\mathrm{mp}(\mathbf{Z}); \mathrm{ap}(\mathbf{Z})]\right), \tag{1}$$

where $\mathrm{mp}(\cdot)$ and $\mathrm{ap}(\cdot)$ denote max pooling and average pooling along the channel dimension, respectively, $*$ represents the convolution operation, $\mathbf{W}_1$ is the convolution filter weight, and $\sigma(\cdot)$ denotes the sigmoid activation.

After obtaining the clothing-related spatial weight mask $\mathbf{m}_c$, the next step is to guide the identity branch to emphasize clothing-irrelevant regions while suppressing clothing-related ones. A straightforward solution would be to directly invert $\mathbf{m}_c$ to construct the identity attention map. However, such a hard exclusion may inevitably remove implicit identity cues that are still present within clothing regions, thereby weakening the model's discriminative ability.

Consequently, we introduce a smooth suppression strategy that adaptively balances the influence of clothing areas through a learnable coefficient λ. Specifically, we generate the identity-related spatial weight mask $\mathbf{m}_{id}$ as:

$$\mathbf{m}_{id} = \mathbf{1} - \lambda \cdot \mathbf{m}_c, \tag{2}$$

where $\mathbf{1}$ denotes an all-one matrix with the same dimensions as $\mathbf{m}_c$, λ is a learnable parameter that controls the degree of suppression on clothing regions. This design allows the network to retain a small proportion of potentially informative identity details within clothing areas while effectively reducing their interference.

After obtaining the spatial weight masks $\mathbf{m}_c$ and $\mathbf{m}_{id}$, we can obtain the disentangled identity features $\mathbf{f}$ and clothing-related features $\mathbf{f}^c$, which can be formulated as:

$$\mathbf{f} = \mathrm{P}(\mathbf{m}_{id} \odot \mathbf{Z}), \quad \mathbf{f}^c = \mathrm{P}(\mathbf{m}_c \odot \mathbf{Z}), \tag{3}$$

where $\odot$ denotes element-wise multiplication, $\mathrm{P}(\cdot)$ denotes the global pooling operation.

Each branch is supervised by an individual classification objective to ensure discriminative representation learning for both identity and clothing. Specifically, cross-entropy loss functions are applied to the identity and clothing features, respectively:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i^{id}|\mathbf{f}_i;\theta_{id}) - \frac{1}{N}\sum_{i=1}^{N} \log p(y_i^{c}|\mathbf{f}_i^{c};\theta_{c}), \tag{4}$$

where N denotes the batch size, $p(\cdot|\cdot;\theta)$ denotes the predicted class probability parameterized by the corresponding classifier head weights θ_{id} and θ_c for identity and clothing branches, respectively.

To further encourage the semantic disentanglement between identity and clothing representations, we introduce an orthogonal constraint loss that minimizes the correlation between the two feature spaces:

$$\mathcal{L}_{orth} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\langle \mathbf{f}_i, \mathbf{f}_i^c \rangle|}{\|\mathbf{f}_i\|_2 \cdot \|\mathbf{f}_i^c\|_2}, \tag{5}$$

where $|\langle \cdot, \cdot \rangle|$ represents the absolute value after the dot product of two feature embeddings. This constraint enforces the two subspaces to be decorrelated, allowing the network to better capture identity-consistent semantics while suppressing clothing-related interference.

4.2 Bi-Directional Prototype Learning

After obtaining the disentangled identity features from the DBDL module, the next challenge is to achieve cross-modality alignment while further preventing clothing interference. To this end, we propose a BPL module to achieve both intra-modality compactness and inter-modality consistency within the same identity. Specifically, BPL simultaneously aggregates identity features within each modality and aligns each sample with its corresponding prototype in the other modality, effectively reinforcing identity consistency under dual heterogeneity.

Notation and Preliminaries. To facilitate the description of our method, the disentangled identity features $\mathbf{f}$ extracted from a training batch are divided into a visible feature set $F^V = \{\mathbf{f}_1^V, \ldots, \mathbf{f}_M^V\}$ and an infrared feature set $F^I = \{\mathbf{f}_1^I, \ldots, \mathbf{f}_M^I\}$, where $\mathbf{f}_i^V$ and $\mathbf{f}_i^I$ denote the identity representations of the i-th visible and infrared instances, respectively, and M is the number of instances in each modality. During training, a balanced sampling strategy is employed to ensure that each identity appears with T instances per modality in every mini-batch.

Furthermore, we maintain two modality-specific prototype sets for the visible and infrared modalities:

$$P^V = \{\mathbf{p}_1^V, \ldots, \mathbf{p}_K^V\}, \quad P^I = \{\mathbf{p}_1^I, \ldots, \mathbf{p}_K^I\}, \tag{6}$$

where $\mathbf{p}_k^V$ and $\mathbf{p}_k^I$ denote the prototypes corresponding to the k-th identity in the visible and infrared modalities, and K is the total number of identities in the training set.

Modality-specific Prototype Initialization. Each prototype is initialized by averaging the identity features of all samples belonging to the same identity within the corresponding modality:

$$\mathbf{p}_k^V = \frac{1}{T}\sum_{y_i^{id}=k}\mathbf{f}_i^V, \quad \mathbf{p}_k^I = \frac{1}{T}\sum_{y_i^{id}=k}\mathbf{f}_i^I, \tag{7}$$

where T is the number of IR/RGB instances per identity in a batch, y_i^{id} is the identity label of the i-th instance.

Momentum Updating. During training, the prototypes are dynamically updated each iteration following a momentum-based strategy. Given the mini-batch mean feature of identity k in modality $m \in \{V, I\}$, denoted as $\bar{\mathbf{f}}_k^m$ (computed from samples in the current batch with $y_i^{id} = k$), the update rule is defined as:

$$\mathbf{p}_k^V(\delta) \leftarrow \alpha\, \mathbf{p}_k^V(\delta - 1) + (1 - \alpha)\, \bar{\mathbf{f}}_k^V, \tag{8}$$

$$\mathbf{p}_k^I(\delta) \leftarrow \alpha\, \mathbf{p}_k^I(\delta - 1) + (1 - \alpha)\, \bar{\mathbf{f}}_k^I, \tag{9}$$

where α is the momentum coefficient and δ denotes the current training iteration. This momentum mechanism stabilizes the prototype evolution by suppressing batch-level noise and enables each prototype to gradually approximate the global identity representation, leading to smoother convergence and improved cross-modality generalization.

Intra-modality Prototype Learning. Within the same modality, large intra-class variations may still exist due to clothing changes, pose variations, and etc, which can undermine the compactness of identity representations. To alleviate this issue, we encourage each sample to be pulled closer to its corresponding identity prototype while being pushed away from other prototypes within the same modality. Formally, the intra-modality prototype loss is formulated in a ProtoNCE [24] manner as:

$$\mathcal{L}_{intra}^V = -\frac{1}{M} \sum_{i=1}^{M} \log \frac{\exp\left(\mathbf{f}_i^V \cdot \mathbf{p}_{y_i^{id}}^V / \tau\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{f}_i^V \cdot \mathbf{p}_k^V / \tau\right)}, \tag{10}$$

$$\mathcal{L}_{intra}^I = -\frac{1}{M} \sum_{i=1}^{M} \log \frac{\exp\left(\mathbf{f}_i^I \cdot \mathbf{p}_{y_i^{id}}^I / \tau\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{f}_i^I \cdot \mathbf{p}_k^I / \tau\right)}, \tag{11}$$

where τ is a temperature hyperparameter, which is fixed to 1/16 following [20].

Inter-modality Prototype Learning. While intra-modality prototype learning effectively compacts features within each modality, achieving intra-modality consistency, the next goal is to align identity semantics across modalities.

To ensure identity discrimination in cross-modality retrieval, a visible sample should be closer to its corresponding identity prototype in the infrared space than to any other identity prototypes, and vice versa. Motivated by this intuition, we introduce inter-modality prototype learning, which enforces modality-invariant identity alignment. This mechanism reduces the large appearance discrepancy caused by the modality gap and encourages the model to focus on discriminative identity-related clues. Formally, the inter-modality prototype loss is defined as:

$$\mathcal{L}_{inter}^V = -\frac{1}{M} \sum_{i=1}^{M} \log \frac{\exp\left(\mathbf{f}_i^V \cdot \mathbf{p}_{y_i^{id}}^I / \tau\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{f}_i^V \cdot \mathbf{p}_k^I / \tau\right)}, \tag{12}$$

$$\mathcal{L}_{inter}^I = -\frac{1}{M} \sum_{i=1}^{M} \log \frac{\exp\left(\mathbf{f}_i^I \cdot \mathbf{p}_{y_i^{id}}^V / \tau\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{f}_i^I \cdot \mathbf{p}_k^V / \tau\right)}. \tag{13}$$

4.3 Progressive Learning

Stage I: Optimize with DBDL. The first stage aims to purify identity representations before any cross-modality alignment. Since aligning noisy features may reinforce clothing-induced biases, we optimize only the DBDL module in this stage to extract clothing-invariant and modality-robust identity cues. The training objective is defined as:

$$\mathcal{L}_{\mathrm{I}} = \mathcal{L}_{base} + \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{orth}, \tag{14}$$

where λ_1 regulates the strength of the orthogonality constraint for effective feature disentanglement.

Stage II: Jointly optimize with BPL. Once disentangled identity features are obtained, we introduce cross-modality alignment on top of these purified representations to mitigate modality gap without reintroducing clothing interference. In this stage, the DBDL and BPL modules are jointly optimized: BPL further suppresses clothing-induced noise via intra-modality prototype refinement, while simultaneously bridging spectral discrepancy through inter-modality prototype alignment. The training loss in Stage II is:

$$\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\mathrm{I}} + \mathcal{L}^{V}_{intra} + \mathcal{L}^{I}_{intra} + \lambda_2(\mathcal{L}^{V}_{inter} + \mathcal{L}^{I}_{inter}), \tag{15}$$

where λ_2 controls the contribution of intra-modality prototype learning.

5 Experiments

We conduct comprehensive experiments to evaluate the effectiveness of PIA and compare it with state-of-the-art methods on both VI-ReID and CC-ReID tasks. All evaluations are performed on the SYSU-CMCC dataset. For quantitative assessment, we adopt the Cumulative Matching Characteristic (CMC) curve and the mean Average Precision (mAP) as evaluation metrics. Following standard protocols, we report the results under V2I and I2V modes.

5.1 Implementation Details

All experiments are conducted on a single NVIDIA A100 GPU. For our proposed PIA, we adopt ResNet-50 [12] pre-trained on ImageNet [5] as the backbone. Following [34], we remove the last spatial downsampling operation to enhance feature granularity. In line with [15], we apply global average pooling and global max pooling to the backbone's output feature map, concatenate the pooled features, and normalize the final feature representation using Batch Normalization [17]. All input images are resized to 384×192. Data augmentation includes random horizontal flipping, random cropping, and random erasing [70]. The model is trained via Adam optimizer [22] for 90 epochs and the Stage II training loss $\mathcal{L}_{\mathrm{II}}$ is used after 55-th epoch. The learning rate is initialized at 3.5×10^{-4} and decayed by a factor of 10 every 30 epochs. Each mini-batch contains 64 images,

Table 2: Comparison with state-of-the-art VI-ReID and CC-ReID methods on the SYSU-CMCC dataset. The best and second-best results are highlighted in bold and underlined.

Type	Method	SYSU-CMCC							
		Visible to Infrared				Infrared to Visible			
		R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
VI-ReID	DDAG [58]	18.3	52.1	54.7	19.6	16.8	49.7	54.6	17.2
	CAJ [57]	21.5	53.4	56.9	20.7	19.8	50.6	56.2	19.6
	MPANet [49]	19.8	51.2	55.5	19.9	18.4	51.1	57.8	19.0
	PMT [32]	18.6	50.8	57.4	19.2	17.9	52.5	58.7	18.9
	SAAI [6]	20.4	52.3	57.6	21.2	19.1	50.2	58.6	19.2
	MCLNet [11]	22.6	56.2	60.3	23.4	21.4	56.4	59.4	21.7
	HOS-Net [38]	23.1	57.5	59.6	24.3	21.7	57.2	61.4	22.7
	MMN [67]	21.8	54.2	56.1	21.3	20.5	55.6	59.3	19.7
	IDKL [39]	22.4	56.0	60.3	21.7	21.3	55.7	60.8	22.3
	DEEN [66]	28.3	56.9	67.7	24.8	24.7	52.4	60.5	28.5
CC-ReID	GI-ReID [19]	24.5	53.4	62.1	22.6	21.3	49.7	57.8	23.1
	AIM [56]	37.7	58.0	67.2	28.8	28.7	43.4	49.5	28.4
	CAL [8]	40.1	63.2	76.2	33.2	35.6	<u>58.2</u>	65.3	37.1
	CSCI [36]	<u>42.9</u>	60.8	73.1	31.9	37.6	56.7	64.8	36.9
	CaAug [28]	42.6	<u>65.6</u>	<u>77.4</u>	<u>36.6</u>	<u>38.2</u>	57.4	<u>66.4</u>	<u>38.2</u>
Our	PIA	**57.0**	**78.6**	**85.3**	**46.5**	**50.4**	**65.3**	**69.6**	**50.3**

corresponding to 8 identities, where each identity contributes 4 visible and 4 infrared samples to ensure balanced cross-modality training.

Hyperparameters. The momentum coefficient α in Eq. (8) and Eq. (9) is fixed at 0.9, while the balancing weights λ_1 and λ_2 in Eq. (14) and Eq. (15) are set to 0.5 and 1.5, respectively. All hyperparameters are selected through grid search.

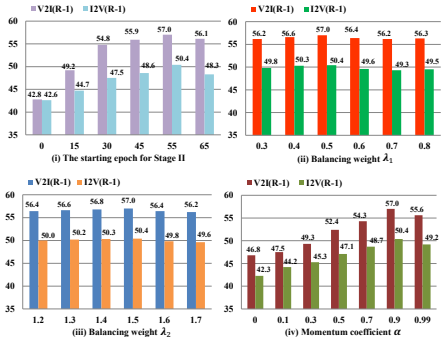
5.2 Performance Comparisons and Analysis

Comparison with State-of-the-Art Methods. We report the performance comparison between PIA and existing state-of-the-art VI-ReID and CC-ReID methods on the SYSU-CMCC dataset. As shown in Tab. 2, PIA significantly surpasses all existing methods under both V2I and I2V settings. Specifically, PIA achieves the best **57.0%**/**46.5%**, **50.4%**/**50.3%** Rank-1/mAP accuracy on V2I setting and I2V setting, respectively, exceeding the best-performing VI-ReID methods by a large margin. Compared with the strongest CC-ReID approaches, PIA still exhibits clear advantages, demonstrating its superior ability to handle modality discrepancy and appearance variation simultaneously in CMCC-ReID.

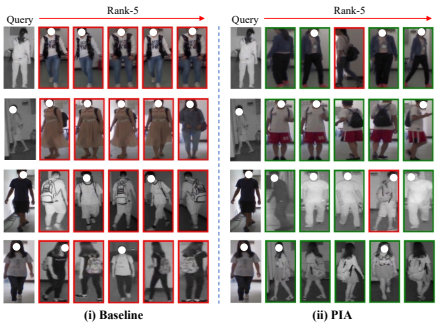
Comparison between VI-ReID methods and CC-ReID methods. Another remarkable observation is that VI-ReID methods perform significantly worse than CC-ReID ones. This result further validates our earlier finding that modality alignment alone is insufficient to maintain identity consistency under CMCC-ReID, highlighting the necessity of the proposed progressive learning

Table 3: Ablation studies of each component over the SYSU-CMCC dataset.

Index	Baseline	DBDL	$\mathcal{L}_{orth}$	Progressive	BPL		V2I		I2V	
					$\mathcal{L}_{intra}$	$\mathcal{L}_{inter}$	R-1	mAP	R-1	mAP
1	✔	✗	✗	✗	✗	✗	40.1	33.2	35.6	37.1
2	✔	✔	✗	✗	✗	✗	43.2	35.7	38.6	38.7
3	✔	✔	✔	✗	✗	✗	45.6	37.6	40.5	39.1
4	✔	✔	✔	✔	✔	✗	46.9	39.5	42.7	41.6
5	✔	✔	✔	✔	✔	✔	**57.0**	**46.5**	**50.4**	**50.3**
6	✔	✔	✔	✗	✔	✔	42.8	35.4	42.6	40.7



(a) The impact of (i) the starting epoch for Stage II, (ii) balancing weight λ_1 in Eq. (14), (iii) λ_2 in Eq. (15), and (iv) momentum coefficient α in Eq. (8).



(b) Comparison of Rank-5 retrieval results. The top two rows correspond to the I2V setting, while the bottom two rows correspond to the V2I setting.

Fig. 5: (a) Results of parameter sensitivity analysis for key hyperparameters in our model. (b) Visualization of retrieval results.

strategy, which suppresses clothing interference before cross-modality alignment to ensure stable identity representation.

5.3 Ablation Studies

We conduct comprehensive ablation studies on the SYSU-CMCC dataset to evaluate the effectiveness of each core component in PIA, including the DBDL module, the orthogonal constraint loss $\mathcal{L}_{orth}$, the intra-modality prototype loss $\mathcal{L}_{intra}$ and inter-modality prototype loss $\mathcal{L}_{inter}$ within the BPL module. A reproduced CAL serves as the baseline, and the detailed results are reported in Tab. 3. Note that all results involving BPL are obtained under the proposed progressive learning scheme.

Effectiveness of DBDL. From Index-1 to Index-2, introducing the DBDL module brings 3.1%/2.5% and 3.0%/1.6% improvement in Rank-1/mAP under the V2I and I2V settings, respectively. This demonstrates that DBDL effectively

preserves identity-consistent semantics while mitigating clothing interference by modeling two disentangled feature streams.

Effectiveness of $\mathcal{L}_{orth}$. Comparing Index-2 and Index-3, adding the orthogonality constraint $\mathcal{L}_{orth}$ further boosts performance. This constraint explicitly enforces the independence between disentangled subspaces, preventing clothing-related cues from leaking into the identity branch.

Effectiveness of $\mathcal{L}_{intra}$. Progressing from Index-3 to Index-4, incorporating $\mathcal{L}_{intra}$ consistently enhances the performance. By compacting identity features within each modality, this loss mitigates intra-modality discrepancies caused by pose, illumination, and clothing variations,resulting in more robust and coherent identity clusters.

Effectiveness of $\mathcal{L}_{inter}$. From Index-4 to Index-5, adding $\mathcal{L}_{inter}$ notably improves Rank-1/mAP by 10.1%/7.0% and 7.7%/8.7% under the V2I and I2V settings, respectively. This indicates that $\mathcal{L}_{inter}$ plays a critical role in bridging the cross-modality gap. By enforcing modality-invariant identity alignment, it establishes consistent identity semantics.

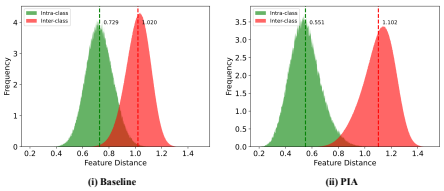
Effectiveness of progressive learning. Comparing Index-6 with Index-3, introducing the BPL without the progressive learning scheme leads to only marginal gains or even performance degradation on certain metrics, highlighting that the effectiveness of PIA primarily stems from the synergistic integration of DBDL, BPL, and the progressive learning framework, all of which are indispensable. To further investigate the impact of the progressive learning paradigm, we vary the epoch at which Stage II training begins and evaluate the corresponding performance, as shown in Fig. 5a(i). The case with 0 epochs corresponds to training without the progressive learning scheme, yielding the lowest accuracy and indicating poor identity-consistent representation under CMCC-ReID. With progressive learning, performance gradually improves as Stage II starts later, achieving the best results when initialized at the 55th epoch, after which performance begins to decline. These findings confirm the effectiveness of the progressive learning strategy.

Hyperparameter Analysis. We further conduct a detailed analysis of the hyperparameters under both the V2I and I2V settings on the SYSU-CMCC dataset. As illustrated in Fig. 5a, the model achieves the optimal performance when λ_1 is set to 0.5 and λ_2 to 1.5. Notably, the overall performance remains relatively stable across a wide range of values, indicating that our PIA is not sensitive to the precise choice of these training balancing weight.

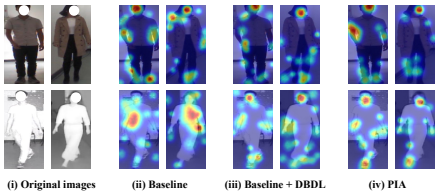
As for the impact of momentum coefficient α in Eq. (8), an excessively small or large value can hinder the stability of prototype representations, leading to notable performance degradation. As illustrated in Fig. 5a(iv), the best results are achieved when α is set to 0.9, indicating that a moderate momentum facilitates more stable and discriminative prototype updates.

5.4 Visualization

Retrieval Results. To qualitatively validate the effectiveness of PIA, Fig. 5b presents Rank-5 retrieval comparisons between PIA and the baseline CAL on



(a) The cosine distance distributions of positive and negative pairs on the SYSU-CMCC dataset during inference.



(b) The visualization of attention maps on SYSU-CMCC. (i) original images, (ii) attention maps generated by the baseline CAL, (iii) attention maps from the baseline with DBDL, and (iv) attention maps from the complete PIA.

Fig. 6: Visualizations of (a) cosine distance distributions and (b) attention maps.

SYSU-CMCC. For each query, correctly retrieved images are marked with green boxes, while incorrect ones are marked in red. Compared with the baseline, PIA consistently returns more correct matches within the top positions, demonstrating its stronger ability to capture identity cues in CMCC-ReID.

Feature Distribution Analysis. Fig. 6a shows the cosine distance distributions of positive and negative pairs on SYSU-CMCC during inference. Compared with the baseline CAL, FID notably decreases the mean distance of positive pairs from 0.729 to 0.551, while increasing that of negative pairs from 1.02 to 1.102.

Attention Maps. We visualize the attention maps of the baseline model, the baseline with DBDL, and the complete PIA in Fig. 6b. The baseline model exhibits noisy attention introduced by background clutter, failing to suppress clothing-related regions under the infrared modality. With the introduction of DBDL, the model effectively reduces attention to clothing areas and shows improved robustness in information-degraded conditions. Nevertheless, some irrelevant activations still persist in background regions. In contrast, PIA further suppresses both clothing-related and background distractions, producing more concentrated and identity-consistent attention across modalities.

6 Conclusion

In this paper, we introduce a new task, termed CMCC-ReID, which aims to match pedestrian images across variations in both modality and clothing, and we establish the SYSU-CMCC dataset, providing the first benchmark for CMCC-ReID. To address CMCC-ReID, we propose a Progressive Identity Alignment Network (PIA) that first disentangles clothing-irrelevant identity features through a novel Dual-Branch Disentanglement Learning (DBDL) module, and then progressively aligns modality representations via the proposed Bi-Directional Prototype Learning (BPL) module. Extensive experiments demonstrate that PIA significantly outperforms existing methods, setting a strong baseline for this challenging task.

References

1. Chen, J., Jiang, X., Wang, F., Zhang, J., Zheng, F., Sun, X., Zheng, W.S.: Learning 3d shape feature for texture-insensitive person re-identification. In: CVPR. pp. 8146–8155 (2021)
2. Chen, Y., Wan, L., Li, Z., Jing, Q., Sun, Z.: Neural feature search for rgb-infrared person re-identification. In: CVPR. pp. 587–597 (2021)
3. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: CVPR. pp. 10257–10266 (2020)
4. Dai, W., Lu, L., Li, Z.: Diffusion-based synthetic data generation for visible-infrared person re-identification. In: AAAI. vol. 39, pp. 11185–11193 (2025)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
6. Fang, X., Yang, Y., Fu, Y.: Visible-infrared person re-identification via semantic alignment and affinity inference. In: ICCV. pp. 11270–11279 (2023)
7. Feng, J., Wu, A., Zheng, W.S.: Shape-erased feature learning for visible-infrared person re-identification. In: CVPR. pp. 22752–22761 (2023)
8. Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., Chen, X.: Clothes-changing person re-identification with rgb modality only. In: CVPR. pp. 1060–1069 (2022)
9. Guo, W., Pan, Z., Liang, Y., Xi, Z., Zhong, Z., Feng, J., Zhou, J.: Lidar-based person re-identification. In: CVPR. pp. 17437–17447 (2024)
10. Han, K., Gong, S., Huang, Y., Wang, L., Tan, T.: Clothing-change feature augmentation for person re-identification. In: CVPR. pp. 22066–22075 (2023)
11. Hao, X., Zhao, S., Ye, M., Shen, J.: Cross-modality person re-identification via modality confusion and center aggregation. In: ICCV. pp. 16403–16412 (2021)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
13. He, W., Deng, Y., Tang, S., Chen, Q., Xie, Q., Wang, Y., Bai, L., Zhu, F., Zhao, R., Ouyang, W., et al.: Instruct-reid: A multi-purpose person re-identification task with instructions. In: CVPR. pp. 17521–17531 (2024)
14. Hong, P., Wu, T., Wu, A., Han, X., Zheng, W.S.: Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: CVPR. pp. 10513–10522 (2021)
15. Huang, Y., Wu, Q., Xu, J., Zhong, Y., Zhang, Z.: Clothing status awareness for long-term person re-identification. In: ICCV. pp. 11895–11904 (2021)
16. Huang, Z., Liu, J., Li, L., Zheng, K., Zha, Z.J.: Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In: AAAI. vol. 36, pp. 1034–1042 (2022)
17. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456 (2015)
18. Jiang, K., Zhang, T., Liu, X., Qian, B., Zhang, Y., Wu, F.: Cross-modality transformer for visible-infrared person re-identification. In: ECCV. pp. 480–496 (2022)
19. Jin, X., He, T., Zheng, K., Yin, Z., Shen, X., Huang, Z., Feng, R., Huang, J., Chen, Z., Hua, X.S.: Cloth-changing person re-identification from a single image with gait prediction and regularization. In: CVPR. pp. 14278–14287 (2022)
20. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. NeurIPS **33**, 18661–18673 (2020)

21. Kim, M., Kim, S., Park, J., Park, S., Sohn, K.: Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In: CVPR. pp. 18621–18632 (2023)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: AAAI. vol. 34, pp. 4610–4617 (2020)
24. Li, J., Zhou, P., Xiong, C., Hoi, S.C.: Prototypical contrastive learning of unsupervised representations. arXiv preprint arXiv:2005.04966 (2020)
25. Liang, T., Jin, Y., Liu, W., Li, Y.: Cross-modality transformer with modality mining for visible-infrared person re-identification. IEEE TMM **25**, 8432–8444 (2023)
26. Liang, X., Rawat, Y.S.: Differ: Disentangling identity features via semantic cues for clothes-changing person re-id. In: CVPR. pp. 13980–13989 (2025)
27. Liu, F., Ye, M., Du, B.: Dual level adaptive weighting for cloth-changing person re-identification. IEEE TIP **32**, 5075–5086 (2023)
28. Liu, F., Ye, M., Du, B.: Cloth-aware augmentation for cloth-generalized person re-identification. In: ACM MM. pp. 4053–4062 (2024)
29. Liu, F., Kim, M., Gu, Z., Jain, A., Liu, X.: Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In: ICCV. pp. 19617–19626 (2023)
30. Liu, F., Kim, M., Ren, Z., Liu, X.: Distilling clip with dual guidance for learning discriminative human body shape representation. In: CVPR. pp. 256–266 (2024)
31. Liu, J., Sun, Y., Zhu, F., Pei, H., Yang, Y., Li, W.: Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In: CVPR. pp. 19366–19375 (2022)
32. Lu, H., Zou, X., Zhang, P.: Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In: AAAI. vol. 37, pp. 1835–1843 (2023)
33. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: CVPR. pp. 13379–13389 (2020)
34. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPRW. pp. 0–0 (2019)
35. 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)
36. Pathak, P., Rawat, Y.S.: Colors see colors ignore: Clothes changing reid with color disentanglement. In: ICCV. pp. 16797–16807 (2025)
37. Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.G., Xue, X.: Long-term cloth-changing person re-identification. In: ACCV (2020)
38. 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: AAAI. vol. 38, pp. 4596–4604 (2024)
39. Ren, K., Zhang, L.: Implicit discriminative knowledge learning for visible-infrared person re-identification. In: CVPR. pp. 393–402 (2024)
40. Wan, F., Wu, Y., Qian, X., Chen, Y., Fu, Y.: When person re-identification meets changing clothes. In: CVPRW. pp. 830–831 (2020)
41. Wang, G.A., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X., Hou, Z.G.: Cross-modality paired-images generation for rgb-infrared person re-identification. In: AAAI. vol. 34, pp. 12144–12151 (2020)

42. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: ICCV. pp. 3623–3632 (2019)
43. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: CVPR. pp. 618–626 (2019)
44. Wei, X., Song, K., Yang, W., Yan, Y., Meng, Q.: A visible-infrared clothes-changing dataset for person re-identification in natural scene. Neurocomputing **569**, 127110 (2024)
45. Wei, Z., Yang, X., Wang, N., Gao, X.: Syncretic modality collaborative learning for visible infrared person re-identification. In: ICCV. pp. 225–234 (2021)
46. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: ECCV. pp. 3–19 (2018)
47. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: ICCV. pp. 5380–5389 (2017)
48. Wu, J., Liu, H., Su, Y., Shi, W., Tang, H.: Learning concordant attention via target-aware alignment for visible-infrared person re-identification. In: ICCV. pp. 11122–11131 (2023)
49. 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: CVPR. pp. 4330–4339 (2021)
50. Wu, Y., Meng, L.C., Zichao, Y., Chan, S., Wang, H.Q.: Wrim-net: Wide-ranging information mining network for visible-infrared person re-identification. In: ECCV. pp. 55–72 (2024)
51. Xu, H., Li, B., Niu, G.: Identity-aware feature decoupling learning for clothing-change person re-identification. In: ICASSP. pp. 1–5 (2025)
52. Xu, H., Niu, G.: Bit: Matching-based bi-directional interaction transformation network for visible-infrared person re-identification. In: CVPR. pp. 40386–40396 (2026)
53. Xu, P., Zhu, X.: Deepchange: A long-term person re-identification benchmark with clothes change. In: ICCV. pp. 11196–11205 (2023)
54. Yang, B., Chen, J., Ye, M.: Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In: CVPR. pp. 16870–16879 (2024)
55. Yang, Q., Wu, A., Zheng, W.S.: Person re-identification by contour sketch under moderate clothing change. IEEE TPAMI **43**(6), 2029–2046 (2019)
56. Yang, Z., Lin, M., Zhong, X., Wu, Y., Wang, Z.: Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In: CVPR. pp. 1472–1481 (2023)
57. Ye, M., Ruan, W., Du, B., Shou, M.Z.: Channel augmented joint learning for visible-infrared recognition. In: ICCV. pp. 13567–13576 (2021)
58. Ye, M., Shen, J., J. Crandall, D., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: ECCV. pp. 229–247 (2020)
59. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE TPAMI **44**(6), 2872–2893 (2021)
60. Yu, H., Cheng, X., Peng, W., Liu, W., Zhao, G.: Modality unifying network for visible-infrared person re-identification. In: ICCV. pp. 11185–11195 (2023)
61. Yu, S., Li, S., Chen, D., Zhao, R., Yan, J., Qiao, Y.: Cocas: A large-scale clothes changing person dataset for re-identification. In: CVPR. pp. 3400–3409 (2020)

62. Yuan, C., Liu, Z., Zhang, G., Xu, H., Zhao, Y., Niu, G., Li, B.: Modality-transition representation learning for visible-infrared person re-identification. arXiv preprint arXiv:2511.02685 (2025)
63. Zhang, G., Zhang, Y., Zhang, T., Li, B., Pu, S.: Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In: CVPR. pp. 14133–14142 (2023)
64. Zhang, Q., Lai, C., Liu, J., Huang, N., Han, J.: Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In: CVPR. pp. 7349–7358 (2022)
65. Zhang, Y., Zhao, S., Kang, Y., Shen, J.: Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In: ECCV. pp. 462–479 (2022)
66. Zhang, Y., Wang, H.: Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In: CVPR. pp. 2153–2162 (2023)
67. Zhang, Y., Yan, Y., Lu, Y., Wang, H.: Towards a unified middle modality learning for visible-infrared person re-identification. In: ACM MM. pp. 788–796 (2021)
68. Zhao, Y., Liu, H., Niu, G.: Mos: Mitigating optical-sar modality gap for cross-modal ship re-identification. In: CVPR. pp. 30335–30345 (2026)
69. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV. pp. 1116–1124 (2015)
70. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI. vol. 34, pp. 13001–13008 (2020)