CKDA: Cross-modality Knowledge Disentanglement and Alignment for Visible-Infrared Lifelong Person Re-identification

Zhenyu Cui, Jiahuan Zhou, Yuxin Peng*

Wangxuan Institute of Computer Technology, Peking University
cuizhenyu@stu.pku.edu.cn, {jiahuanzhou, pengyuxin}@pku.edu.cn

Abstract

Lifelong person Re-IDentification (LReID) aims to match the same person employing continuously collected individual data from different scenarios. To achieve continuous all-day person matching across day and night, Visible-Infrared Lifelong person Re-IDentification (VI-LReID) focuses on sequential training on data from visible and infrared modalities and pursues average performance over all data. To this end, existing methods typically exploit cross-modal knowledge distillation to alleviate the catastrophic forgetting of old knowledge. However, these methods ignore the mutual interference of modality-specific knowledge acquisition and modality-common knowledge anti-forgetting, where conflicting knowledge leads to collaborative forgetting. To address the above problems, this paper proposes a Cross-modality Knowledge Disentanglement and Alignment method, called CKDA, which explicitly separates and preserves modality-specific knowledge and modality-common knowledge in a balanced way. Specifically, a Modality-Common Prompting (MCP) module and a Modality-Specific Prompting (MSP) module are proposed to explicitly disentangle and purify discriminative information that coexists and is specific to different modalities, avoiding the mutual interference between both knowledge. In addition, a Cross-modal Knowledge Alignment (CKA) module is designed to further align the disentangled new knowledge with the old one in two mutually independent inter- and intra-modality feature spaces based on dual-modality prototypes in a balanced manner. Extensive experiments on four benchmark datasets verify the effectiveness and superiority of our CKDA against state-of-the-art methods. The source code of this paper is available at https://github.com/PKU-ICST-MIPL/CKDA-AAAI2026.

Introduction

Lifelong person Re-IDentification (LReID) aims to employ sequentially collected data to match the same person in all scenarios (Pu et al. 2021, 2023). Its core challenge lies in mitigating catastrophic forgetting of old knowledge learned in known scenarios while continuously acquiring new knowledge from new scenarios (De Lange et al. 2021; Masana et al. 2022; Xu, Zou, and Zhou 2024). To this end, existing LReID methods have demonstrated their effectiveness in suppressing the forgetting of old knowledge through data replay (Yu et al. 2023a; Ge et al. 2022) and knowledge distillation (Sun and Mu 2022; Cui et al. 2024). Despite some progress in visible modality, these methods typically suffer from a more practical and challenging scenario where both visible and infrared images are continuously collected to match a person across day and night, which is also called Visible-Infrared Lifelong person Re-IDentification (VI-LReID) (Xing et al. 2024).

As a preliminary pioneer, Visible-Infrared person Re-IDentification (VI-ReID) matches the same person captured by visible and infrared cameras by extracting discriminative information against the discrepancy of heterogeneous information within different modalities (Ye et al.

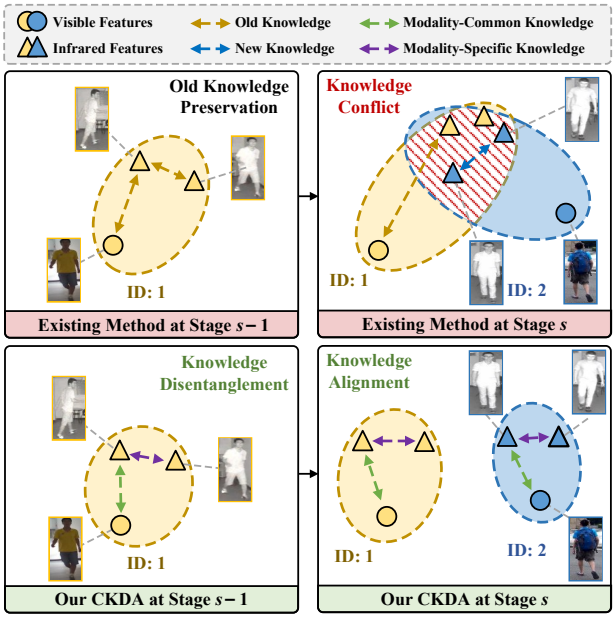


Figure 1: Comparison of different LReID methods when facing visible and infrared images. Existing methods suffer from the conflict between the new knowledge acquisition (*e.g.*, acquiring radiation knowledge specific to the infrared modality) and the old knowledge preservation (*e.g.*, preserving shape knowledge that coexists in both modalities, which conflicts with the former one). While our CKDA achieves knowledge balancing by explicitly aligning the disentangled modality-common and -specific knowledge.

*Corresponding author.

2020; Cui, Zhou, and Peng 2024). To this end, modality-common representation-based methods (Wu et al. 2021; Yu et al. 2023b) extract features from different modalities and map them into a common feature space to align information of the same person, while modality-specific representation-based methods (Ren and Zhang 2024; Yu et al. 2023b) focus on extracting unique information from each modality to further enrich the diversity and distinctiveness of individual representations. Therefore, some recent works attempt to transfer the above methods to alleviate the knowledge forgetting of both modalities in VI-LReID (Xing et al. 2024; Luo et al. 2025). Among them, some existing methods follow the data rehearsal paradigm and retain old knowledge by inference task matching (Xing et al. 2024) and distilling cross-modal knowledge (Luo et al. 2025) between old visible and infrared data. Inspired by prompting technology, the recent VI-LReID method (Zhu et al. 2025) introduced a dynamic prompt selection mechanism to preserve instance-level old knowledge while learning the most relevant new knowledge for the current scenario from a modality-common prompt pool.

However, in addition to the privacy issue caused by the data rehearsal (Ahmad et al. 2022), these methods typically ignore the conflict between modality-specific knowledge acquisition and modality-common knowledge preservation, making it difficult to balance the cross-modal knowledge. Specifically, as shown in Figure 1, when learning on new data, existing VI-LReID methods inevitably sacrifice the old knowledge that coexists between modalities (*e.g.*, body shape knowledge) due to accumulating new knowledge that are specific to distinct modality (*e.g.*, thermal radiation knowledge in infrared modality, which is useless or even harmful for matching thermal radiation information that does not exist in the visible modality). As a result, the conflicting modality-specific knowledge acquisition and modality-common knowledge preservation typically lead to joint forgetting due to their mutual interference, which further eliminates the discriminability of cross-modality knowledge in both new and old scenarios.

To tackle the above issue, this paper proposes a Cross-modality Knowledge Disentanglement and Alignment (CKDA) method for the VI-LReID task, which aims to balance the conflicting cross-modality knowledge acquisition and preservation. Specifically, to alleviate the mutual interference between cross-modal knowledge, we first proposed a Modality-Common Prompting (MCP) module and a Modality-Specific Prompting (MSP) module to disentangle both knowledge. The former aims to extract modality-common knowledge that coexists in both modalities, while the latter further purifies knowledge that only exists in visible or infrared modality in a complementary manner. In addition, a Cross-modality Knowledge Aligning (CKA) module is further designed to align the above knowledge with the constructed inter- and intra-modality feature space by employing identity-level prototypes of both modalities, avoiding the rehearsing of old data. In summary, the main contributions of this paper are as follows:

- A Cross-modality Knowledge Disentanglement and Alignment (CKDA) method is proposed to tackle the Visible-Infrared Lifelong person Re-IDentification challenge, which alleviates catastrophic forgetting derived from the conflict between cross-modality knowledge acquisition and preservation.
- A Modality-Common Prompting (MCP) module and a Modality-Specific Prompting (MSP) module are designed to disentangle and purify modality-common and -specific knowledge, explicitly suppressing the mutual interference between both modalities.
- A Cross-modal Knowledge Alignment (CKA) module is designed to balance the disentangled knowledge with a pair of symmetric inter- and intra-modality feature spaces constructed by visible and infrared identity prototypes.

Related Work

Lifelong Person Re-identification

Other than ReID in pre-collected data (Cui et al. 2023), Lifelong person Re-IDentification (LReID) aims to employ sequentially collected data to match the same person in all scenarios (Pu et al. 2021). To this end, replay-based LReID methods rehearse old data when learning on new data to tackle the catastrophic forgetting challenge (Wu and Gong 2021; Ge et al. 2022; Yu et al. 2023b; Chen, Lagadec, and Bremond 2022), which raises concerns about the data privacy issue. In contrast, non-replay methods are dedicated to distilling from the old model to preserve old discriminative knowledge (Xu, Zou, and Zhou 2024; Cui et al. 2024; Xu et al. 2025a). In addition to the above works that rely on well-labelled data, some recent efforts have also explored lifelong person re-identification in more complex open scenarios, *e.g.*, noisy (Xu et al. 2024a), semi-supervised (Xu et al. 2025c) and unsupervised scenarios (Chen, Lagadec, and Bremond 2022), further expanding its real significance.

However, existing LReID methods typically focus on knowledge balancing within visible data, ignoring the requirement of all-day pedestrian matching across day and night with images captured by visible and infrared cameras, respectively, which limits their applicability in practice.

Visible-Infrared Person Re-identification

Visible-Infrared Person Re-IDentification (VI-ReID) aims to match the same person across images captured by both infrared cameras and visible cameras (Wu et al. 2017). To this end, modality-common representation-based methods (Cui, Zhou, and Peng 2024; Liang et al. 2023; Lu, Zou, and Zhang 2023; Ye et al. 2020; Sun and Zhao 2023; Wu et al. 2021; Park et al. 2021) extract and align common discriminative information that coexists in different modalities, while modality-specific representation-based methods (Ren and Zhang 2024; Yu et al. 2023b; Li et al. 2022; Hu et al. 2022; Zhang et al. 2022; Jiang et al. 2022) focus on exploiting and leveraging information that only exists in a specific modality, achieving more comprehensive individual information representation.

Despite some progress, these methods typically suffer from severe performance degradation when new data is continuously learned in a streaming manner. Inspired by recent

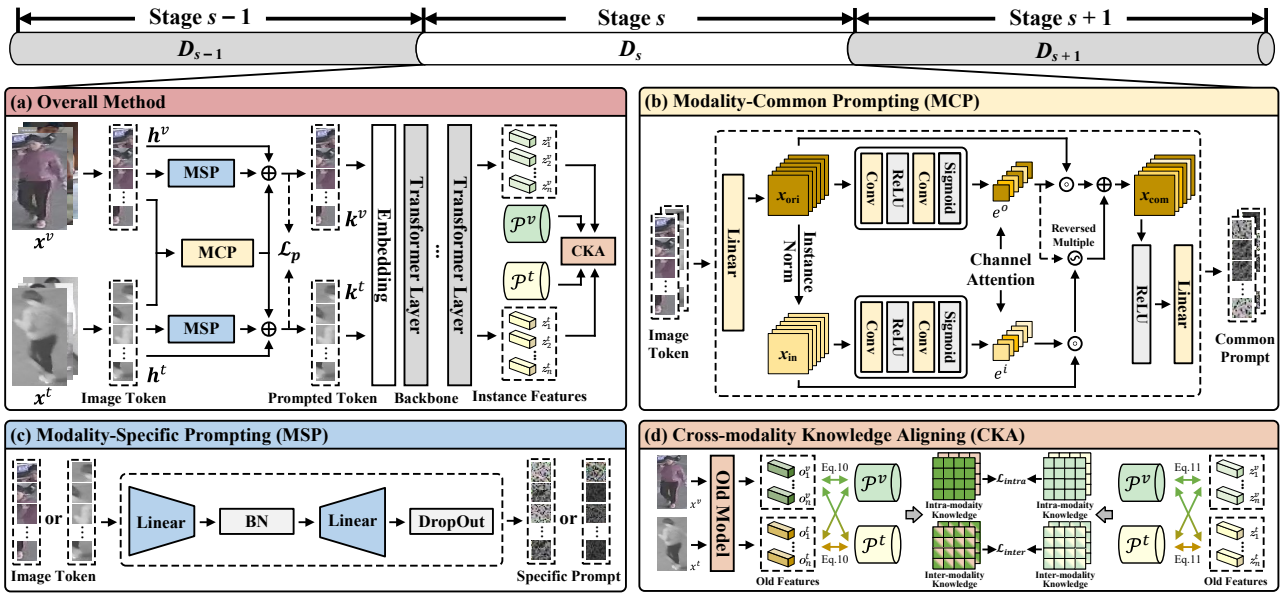


Figure 2: Overview of our proposed Cross-modality Knowledge Disentanglement and Alignment (CKDA) method. The input images are first fed into the Modality-Common Prompting (MCP) module and the Modality-Specific Prompting (MSP) module to generate the corresponding prompted image tokens. Then, the Cross-modality Knowledge Aligning (CKA) module exploits the cross-modality knowledge prototype to align the above two kinds of knowledge, respectively.

advancements in VI-ReID, some works (Xing et al. 2024; Luo et al. 2025) began to focus on a practical and challenging task, called Visible-Infrared Lifelong Person ReIDentification (VI-LReID), which aims to avoid the catastrophic forgetting of cross-modality knowledge when learning from streaming data sequentially collected by visible and infrared cameras. Among them, TTQK (Xing et al. 2024) learns a specific token for each domain, and matches the most appropriate inference token for each query sample to preserve old knowledge. In order to avoid the additional query token selection overhead, DMM (Luo et al. 2025) alleviate the catastrophic forgetting problem by distilling the similarity between samples of different modalities. Inspired by recent advances in visual prompting (Yao et al. 2025; Zhou et al. 2025; Ai et al. 2025b,a), PP-IPG (Zhu et al. 2025) further exploited fine-grained information of each modality at the instance level through a prompt pool. However, in addition to the privacy issue derived from the heavy reliance on old data, these methods typically suffer from the mutual interference of modality-specific knowledge acquisition and modality-common knowledge anti-forgetting, which limits their capacity to balance the cross-modal knowledge.

Different from these methods, we propose a Cross-modality Knowledge Disentanglement and Alignment (CKDA) method for VI-LReID, which avoids mutual interference between different knowledge by explicitly disentangling and adaptive balancing cross-modal knowledge through dual knowledge alignment.

Method

Preliminary

This paper focuses on Visible-Infrared Lifelong person ReIDentification (VI-LReID) task. Formally, given a sequentially collected data stream $D = \{D_1, D_2, ..., \mathcal{D}_S\}$ with S stages, where each dataset $\mathcal{D}_s$ consists of a visible image set $\mathcal{D}_s^v$ and an infrared image set $\mathcal{D}_s^t$. When sequentially learning on $\mathcal{D}_s$, the previous $s-1$ datasets are inapplicable due to the data privacy issue. For inference, the overall re-identification performance of each dataset between different modalities is employed for evaluation.

Overview of CKDA

The pipeline of our CKDA is shown in Figure 2, which consists of a Modality-Common Prompting (MCP) module, a Modality-Specific Prompting (MSP) module and a Cross-modality Knowledge Aligning (CKA) module. Then, we will elaborate on them in the following sections.

Baseline. Inspired by existing LReID methods (Xu, Zou, and Zhou 2024; Xu et al. 2024b), given a batch of input images with identities $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^B \in \mathcal{D}_s$ at stage s, we input each $\boldsymbol{x}_i$ into a backbone network $\phi_s(\cdot)$ with a batch normalization (Ioffe and Szegedy 2015) layer to extract the deep feature $\boldsymbol{z}_i$. Then, an identity classifier $\psi_s(\cdot)$ is exploit to predict the identity probability $\boldsymbol{y}_i'$ for each image $\boldsymbol{x}_i$.

To facilitate the acquisition of new knowledge, we introduce a classification loss $\mathcal{L}_{ce}$ (Luo et al. 2019) and a triplet loss $\mathcal{L}_{trip}$ (Hermans, Beyer, and Leibe 2017) to improve the discriminability of each person. Therefore, the overall optimization objective of our baseline is formulated as:

$$\mathcal{L}_{base} = \mathcal{L}_{ce} + \mathcal{L}_{trip}. \tag{1}$$

Besides, we adopt the Exponential Moving Average (EMA) strategy, which is widely used in existing LReID methods (Xu et al. 2024b, 2025a) to initially avoid the forgetting of old knowledge caused by drastic parameter changes after each training stage as follows:

$$\phi_s = \lambda \cdot \phi_{s-1} + (1-\lambda) \cdot \phi_s^*, \tag{2}$$

where ϕ_s^* denotes the backbone trained after stage s, λ denotes the balance factor between old and new knowledge.

Modality-Common Prompting

To explicitly disentangle the modality-common knowledge, we design a Modality-Common Prompting (MCP) module to capture the discriminative information that co-exists in visible and infrared modalities. Therefore, MCP adaptively incorporates the features after modality discrepancies elimination to purify modality-common information.

Specifically, given an input image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$, we first divide it into M image tokens $\boldsymbol{h} = \{h_i\}_{i=1}^{M} \in \mathbb{R}^{ps\times ps\times pc}$ of the same size, where H, W are the height and width of image $\boldsymbol{x}$, ps is the size of each image token, C and pc is the channel number of $\boldsymbol{x}$ and $\boldsymbol{h}$. Each image token is then embedded into a d-dimension latent feature as follows:

$$\boldsymbol{h}' = \mathcal{E}_d(\boldsymbol{h}), \tag{3}$$

where $\mathcal{E}_d(\cdot)$ denotes the embedding layer of the image token. Then, we rearrange $\boldsymbol{h}'$ to form a feature map $\boldsymbol{x}_{ori} \in \mathbb{R}^{H\times W\times C}$ for knowledge purification.

To purify the modality-common knowledge, we exploit the Instance Normalization (Ulyanov, Vedaldi, and Lempitsky 2016) to alleviate the style discrepancy between different modalities, while avoiding the derived loss of discriminative information. Specifically, We first obtain the normalized features $\boldsymbol{x}_{in}$ as follows:

$$\boldsymbol{x}_{in} = \frac{\boldsymbol{x}_{ori} - \mathrm{E}[\boldsymbol{x}_{ori}]}{\sqrt{\mathrm{Var}[\boldsymbol{x}_{ori}] + \epsilon}}, \tag{4}$$

where the parameter ϵ is applied to avoid division by zero errors, $\mathrm{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ denote the channel-wise mean and the variance, respectively. Then, we evaluate the modality-common knowledge distribution of $\boldsymbol{x}_{ori}$ and $\boldsymbol{x}_{in}$ by generating two channel-wise masks e^o and e^i as follows:

$$\begin{cases} e^o = \sigma(W_2^o \cdot \delta(W_1^o \cdot \boldsymbol{x}_{ori})) \\ e^i = \sigma(W_2^i \cdot \delta(W_1^i \cdot \boldsymbol{x}_{in})) \end{cases}, \tag{5}$$

where $\sigma(\cdot)$ and $\delta(\cdot)$ denote the ReLU and the sigmoid activation function, $W_1^o, W_1^i \in \mathbb{R}^{\frac{d}{r}\times d}$ and $W_2^o, W_2^i \in \mathbb{R}^{d\times\frac{d}{r}}$ denote two learnable parameters. Considering the balance between complexity and efficiency, the dimension reduction r is set to 4. To further balance the discriminability and discrepancy of modality-common knowledge, MCP adaptively fuses $\boldsymbol{x}_{ori}$ and $\boldsymbol{x}_{in}$ as follows:

$$\boldsymbol{x}_{com} = e^o \cdot \boldsymbol{x}_{ori} + (1 - e^o)(e^i \cdot \boldsymbol{x}_{in}), \tag{6}$$

where e^i denotes the importance of modality-common knowledge after eliminating modality discrepancies, while $(1 - e^o)$ supplements the discriminative information after suppressing the modality discrepancy in a dynamic manner according to the importance of the modality-common knowledge in the original feature $\boldsymbol{x}_{ori}$.

Finally, we align the common prompting $\boldsymbol{x}_com$ with the original feature map to generate the common prompt $\boldsymbol{k}_{com}$ by restoring the input dimension based on the feature token $\boldsymbol{h}''$, which is obtained by token division of $\boldsymbol{x}_{com}$ as follows:

$$\boldsymbol{k}_{com} = \mathcal{E}_{pc}(\delta(\boldsymbol{h}'')), \tag{7}$$

where $\mathcal{E}_{pc}(\cdot)$ denotes the embedding restoration layer. In this way, the modality-common knowledge can be dynamically purified in the unified common prompt, alleviating the mutual interference of both modalities.

Modality-Specific Prompting

Although the Modality-Common Prompting module disentangles and purifies the common knowledge between different modalities, it is still limited by the insufficient preservation of modality-specific knowledge, which further suppresses the balance of cross-modal knowledge. Therefore, we further propose a Modality-Specific Prompting (MSP) module to further disentangle the modality-specific discriminative information that only exists in a distinct modality.

To this end, a pair of light-weight modality-specific prompt generation networks is designed for visible and infrared modalities, respectively. Specifically, given an image token $\boldsymbol{h}^m$ of a specific modality, where $m \in \{v, t\}$ represents the visible or the infrared modality. The modality-specific prompt $\boldsymbol{k}_{spe}$ can be calculated as follows:

$$\boldsymbol{k}_{spe}^m = \xi(W_2^m \cdot \theta(W_1^m \cdot \boldsymbol{h}^m)), \tag{8}$$

where $\xi(\cdot)$ denotes the dropout layer, $W_1^m \in \mathbb{R}^{\frac{d}{r}\times d}$, $W_2^m \in \mathbb{R}^{d\times\frac{d}{r}}$ denote the learnable parameters of the linear layer, $\theta(\cdot)$ denotes the Batch Normalization layer. Consequently, the modality-specific prompting is encouraged to further stimulate the difference between modalities with batch-wise regularization, thereby facilitating the model to extract modality-specific information in distinct modalities under explicit modality discrepancies.

Optimization of Modality-Com and MSP module. To alleviate the catastrophic forgetting of cross-modal knowledge due to the continual knowledge change across visible and infrared data of different scenarios, we align the generated prompting of the current stage (Stage s) with that of the previous stage (Stage $s-1$). Let $\boldsymbol{k}_p^{m,(s)} = \boldsymbol{k}_{spe}^{m,(s)} + \boldsymbol{k}_{com}^{(s)}$ denote the image prompt at Stage s, the optimization can be achieved by minimizing the prompting loss $\mathcal{L}_p$, which can be calculated as follows:

$$\mathcal{L}_p = |\boldsymbol{k}_p^{m,(s)} - \boldsymbol{k}_p^{m,(s-1)}|. \tag{9}$$

Finally, we can get the prompted image $\boldsymbol{k}^m$ by merging the original image $\boldsymbol{x}^m$ and $\boldsymbol{k}_p^m$ at the current stage, which is then fed into the backbone network for feature extraction.

Cross-modality Knowledge Aligning

Although the above modules eliminate the mutual interference between different knowledge by disentanglement, it is still challenging to simultaneously alleviate two conflicting knowledge in a unified feature space. Therefore, we further propose a Cross-modality Knowledge Aligning (CKA) module, which aims to align modality-specific knowledge in an intra-modality feature space while aligning modality common knowledge in a dual inter-modality feature space.

To this end, we employ the feature prototype as the basis of the old feature space to measure and align cross-modal knowledge. Specifically, given a pair of visible prototypes $\mathcal{P}_{s-1}^v = \{p_i^v\}_{i=1}^{N_{s-1}}$ and infrared prototypes $\mathcal{P}_{s-1}^t =$

	Method	Publication	Backbone	RegDB		SYSU-MM01		LLCM		VCM		Average	
				mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
	JointTrain	-	ViT-B/16	74.8	79.4	55.8	59.4	60.2	56.3	19.1	33.4	52.5	57.1
	SFT	-	ViT-B/16	6.8	5.7	21.0	21.5	26.4	21.3	17.1	29.0	17.8	19.4
Replay	LwF (Li and Hoiem 2017)	*T-PAMI*	ViT-B/16	26.9	26.1	30.9	30.3	39.1	34.5	18.2	30.0	28.8	30.2
Replay	iCaRL (Rebuffi et al. 2017)	*CVPR*	ViT-B/16	42.4	41.3	32.4	32.4	41.9	36.8	18.5	30.7	33.8	35.3
Replay	BiC (Wu et al. 2019)	*CVPR*	ViT-B/16	40.9	39.3	22.8	21.9	43.3	37.5	16.4	27.4	30.9	31.5
Replay	WA (Zhao et al. 2020)	*CVPR*	ViT-B/16	44.9	43.9	34.6	34.4	44.1	38.7	18.7	29.3	35.6	36.6
Replay	PTKP (Ge et al. 2022)	*AAAI*	ViT-B/16	50.3	47.9	26.6	25.8	43.2	36.5	17.9	27.4	34.5	34.4
Replay	TTQK (Xing et al. 2024)	*ICMR*	ViT-B/16	60.3	58.3	33.6	30.9	45.8	39.6	19.4	32.2	39.8	40.2
Replay	PP-IPG (Luo et al. 2025)	*ICMR*	ViT-B/16	59.0	58.8	32.7	35.2	44.4	50.8	35.3	23.3	42.8	42.0
Non-Replay	LwF (Li and Hoiem 2017)	*T-PAMI*	ResNet	23.4	23.0	9.5	9.7	17.8	13.6	6.8	11.6	14.4	14.6
Non-Replay	AKA (Pu et al. 2021)	*CVPR*	ResNet	26.2	23.9	11.0	10.0	17.2	12.9	7.4	13.0	15.5	15.1
Non-Replay	PatchKD (Sun and Mu 2022)	*MM*	ResNet	52.2	55.2	9.6	7.9	18.5	13.7	3.9	5.9	21.1	20.8
Non-Replay	LSTKC (Xu, Zou, and Zhou 2024)	*AAAI*	ResNet	15.9	13.5	15.1	13.4	22.1	17.4	16.0	26.8	17.3	17.8
Non-Replay	DKP (Xu et al. 2024b)	*CVPR*	ResNet	34.1	33.5	26.4	27.1	32.8	27.9	15.3	26.4	27.2	28.7
Non-Replay	DASK (Xu et al. 2025a)	*AAAI*	ResNet	17.3	15.4	23.6	22.0	29.1	24.1	18.1	25.9	22.0	21.8
Non-Replay	LSTKC++ (Xu et al. 2025b)	*T-PAMI*	ResNet	36.1	38.1	35.7	37.0	29.3	24.9	16.3	26.6	29.4	31.6
Non-Replay	LSTKC (Xu, Zou, and Zhou 2024)	*AAAI*	ViT-B/16	35.0	35.4	32.7	34.3	39.6	35.1	15.1	25.3	**30.6**	**32.5**
Non-Replay	DKP (Xu et al. 2024b)	*CVPR*	ViT-B/16	30.3	32.7	30.9	32.8	37.3	32.2	15.2	26.3	28.4	31.0
Non-Replay	TTQK (Xing et al. 2024)	*CVPR*	ViT-B/16	8.1	6.5	20.8	18.6	19.2	14.9	20.8	32.6	17.2	18.2
Non-Replay	DASK (Xu et al. 2025a)	*AAAI*	ViT-B/16	21.6	25.6	28.1	32.3	39.6	36.2	14.1	27.4	25.9	30.4
Non-Replay	LSTKC++ (Xu et al. 2025b)	*T-PAMI*	ViT-B/16	19.2	19.0	27.1	30.3	32.9	29.2	16.1	29.2	23.8	26.9
Non-Replay	**Ours**	This Paper	ViT-B/16	57.0	60.4	34.7	37.3	37.8	31.5	15.6	28.2	**36.3**	**39.4**

Table 1: Performance Comparison on training order: RegDB→SYSU-MM01→LLCM→VCM.

$\{p_i^t\}_{i=1}^{N_{s-1}}$, where N_{s-1} denote the identity number of the $(s-1)^{th}$ stage, p_i^v and p_i^t denote the visible and infrared feature center of old data extracted by $\phi_{s-1}(\cdot)$, we use the current visible instance features $\{o_i^v\}_{i=1}^{N_s}$ and infrared instance features $\{o_i^t\}_{i=1}^{N_s}$ extracted by $\phi_{s-1}(\cdot)$ to represent the old inter-modality feature space $\boldsymbol{O}$ as follows (taking visible-to-infrared as an example):

$$\boldsymbol{O}_{vt}(i,j) = \frac{\exp(cos(\boldsymbol{o}_i^v, \boldsymbol{p}_j^t)/\tau)}{\sum_{k=1}^{N_{s-1}} \exp(cos(\boldsymbol{o}_i^t, \boldsymbol{p}_k^v)/\tau)}, \quad (10)$$

where $cos(\cdot)$ denotes the cosine similarity, τ is a temperature parameter set as 0.1. Thus, the intra-modality feature space can be represented by $\boldsymbol{O}_{vv}$ and $\boldsymbol{O}_{tt}$, which is employed to measure the modality-common knowledge distribution under the old intra-modality feature space. Similarly, we can further exploit the visible features $\{z_i^v\}_{i=1}^{N_s}$ and infrared features $\{z_i^t\}_{i=1}^{N_s}$ extracted by the current model to represent the new distribution $\boldsymbol{Z}$ after the new knowledge acquisition:

$$\boldsymbol{Z}_{vt}(i,j) = \frac{\exp(cos(\boldsymbol{z}_i^v, \boldsymbol{p}_j^t)/\tau)}{\sum_{k=1}^{N_{s-1}} \exp(cos(\boldsymbol{z}_i^t, \boldsymbol{p}_k^v)/\tau)}. \quad (11)$$

Then, to alleviate the catastrophic forgetting of inter- and intra-modality knowledge, we align the inter- and intra-affinity of each feature with the proposed knowledge balancing loss $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$ as follows:

$$\mathcal{L}_{inter} = \mathcal{L}_{KL}(Y_{v,t}^o || Y_{v,t}^z) + \mathcal{L}_{KL}(Y_{t,v}^o || Y_{t,v}^z), \quad (12)$$

$$\mathcal{L}_{intra} = \mathcal{L}_{KL}(Y_{v,v}^o || Y_{v,v}^z) + \mathcal{L}_{KL}(Y_{t,t}^o || Y_{t,t}^z), \quad (13)$$

$$Y_{v,t}^o = \rho(\boldsymbol{O}_{vt}\boldsymbol{O}_{vt}^\top/\tau), \;\; Y_{v,t}^z = \rho(\boldsymbol{Z}_{vt}\boldsymbol{Z}_{vt}^\top/\tau), \quad (14)$$

where $\mathcal{L}_{KL}$ is the Kullback-Leibler (KL) divergence, and $\rho(\cdot)$ is the softmax function. In summary, the CKA module employs the old prototype as the old feature space to alleviate the cross-modal knowledge shift derived by the acquisition of new knowledge via dual knowledge alignment, thereby balancing the discriminative information between and within visible and infrared modalities.

Overall Optimization

The overall optimization objective for our CDKA can be formalized as follows:

$$\mathcal{L} = \mathcal{L}_{base} + \alpha L_p + \beta(\mu L_{inter} + (1-\mu)\mathcal{L}_{intra}), \quad (15)$$

where α and β denote two hyperparameters when training the model, μ denotes a scaling factor to balance the modality-specific and -common knowledge preservation.

Experiments

Datasets and Evaluation Metrics

Datasets. To verify the effectiveness of our CKDA, we follow the most popular VI-LReID benchmark in (Xing et al. 2024), which consists of 4 widely-used visible-infrared person re-identification datasets, including RegDB (Nguyen et al. 2017), SYSU-MM01 (Wu et al. 2017), LLCM (Zhang and Wang 2023), and HITSZ-VCM (VCM) (Lin et al. 2022). To perform fairly comparison in the real lifelong scenario, we also follow (Xing et al. 2024) and sequentially conduct the above datasets as the training order: RegDB→SYSU-MM01→LLCM→VCM. For the video dataset VCM, we follow the frame-level sampling strategy in (Xing et al. 2024) to generate visible and infrared images.

Evaluation Metrics. To compare our CKDA with existing methods, three popular metrics, including Rank-1 accuracy (R1) (Moon and Phillips 2001), mean Average Precision (mAP) (Zheng et al. 2015) with their Average Forgetting (AF) (Chaudhry et al. 2018), are employed to verify the

Method	TTQK	DASK	DKP	LSTKC	LSTKC++	**Ours**
AF(mAP)	36.5	15.1	12.1	8.7	8.5	**5.1**
AF(R1)	35.5	12.1	12.2	7.7	6.9	**4.9**

Table 2: AF comparison to non-replay methods.

Base	MCP	MSP	CKA	mAP	R1
✓	-	-	-	31.8	33.9
✓	✓	-	-	33.4	35.2
✓	✓	✓	-	34.6	37.4
✓	-	-	✓	34.9	37.9
✓	✓	✓	✓	**36.3**	**39.4**

Table 3: Ablation study of each module in CKDA.

overall performance on all datasets. The former two metrics evaluate the ReID performance, while the last metric describes their forgetting rate via the drop value.

Implementation Details

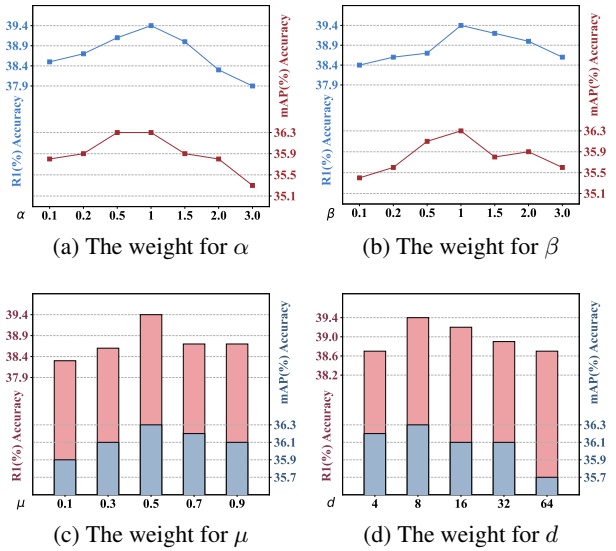
We implement our CKDA on NVIDIA A40 GPU for both training and inference. During the training phase, ViT-B/16 (Dosovitskiy et al. 2020) pre-trained on ImageNet (Russakovsky et al. 2015) is employed as the initial backbone network. For each dataset, we randomly sample 500 identities and train the model for 50 epochs, where inadequate datasets sample all identities. The batch size is set to 64, where each identity has 4 visible images and 4 infrared images. The learning rate is set to 3×10^{-4} and a cosine decay schedule (Chen et al. 2020; Loshchilov and Hutter 2016) to optimize the whole model. During the inference phase, following (Xing et al. 2024), we perform visible-to-infrared re-identification on the RegDB datasets, and perform infrared-to-visible re-identification on the others.

Comparison with State-of-the-arts Methods

We start by compare our proposed CKDA to multiple SOTA methods on both LReID and VI-LReID tasks, including LwF (Li and Hoiem 2017), iCaRL (Rebuffi et al. 2017), BiC (Wu et al. 2019), WA (Zhao et al. 2020), PTKP (Ge et al. 2022), TTQK (Xing et al. 2024), PP-IPG (Luo et al. 2025), AKA (Pu et al. 2021), PatchKD (Sun and Mu 2022), LSTKC (Xu, Zou, and Zhou 2024), DKP (Xu et al. 2024b), DASK (Xu et al. 2025a), and LSTKC++ (Xu et al. 2025b). Note that the best results are marked in **red**, while the sub-optimal results are marked in **blue**.

As shown in Table 1, our CKDA achieves **36.3%/39.4%** on the average mAP/R1 accuracy, which outperforms the existing non-replay SOTA method LSTKC (Xu, Zou, and Zhou 2024) by **5.7%/6.9%**. This is because our CKDA disentangles modality-specific knowledge from modality-common knowledge and thus eliminates their mutual interference. Besides, our method achieves effective dual preservation of both disentangled knowledge by dynamically aligning them in the modality-inter and -intra feature space in a balanced manner, thereby significantly outperforming the state-of-the-art methods.

Besides, Table 2 reports the forgetting degree of our method, which achieves the best average anti-forgetting performance for **5.1%/4.9%** on the average AF(mAP/R1) performance. It implies that our proposed dual modality prompting splits the originally entangled cross-modal knowledge in an explicit way, and further preserves the discriminability of old knowledge through a pair of symmetric inter- and intra-modality distillations. Therefore, our method effectively alleviates the catastrophic forgetting derived from conflicting cross-modal knowledge.



(a) The weight for α (b) The weight for β

(c) The weight for μ (d) The weight for d

Figure 3: The influence of hyperparameters in our CKDA.

Ablation Studies

In this section, we conduct ablation studies to verify the effectiveness of each CKDA component.

Effectiveness of the MCP module. Table 3 reports that MCP module brings **1.6%/1.3%** on average mAP and R1 accuracy compared to base method. It indicates that our approach effectively purifies the discriminative information that coexists in visible and infrared images by suppressing the discrepancy derived from distinct modality styles, thereby improving the adaptability of modality-common knowledge to both modalities.

Effectiveness of the MSP module. Different from the MCP module, the MSP module focuses on extracting discriminative information that only exists in distinct modalities. As shown in Table 3, MSP further improves mAP/R1 by **1.2%/2.2%**, which supports its further separation and complements the modality-specific discriminative information missing from the MCP module. Nevertheless, MSP does not sacrifice the modality-common knowledge introduced by MCP through encouraging distinct discriminative knowledge after amplifying the discrepancies between modalities, thus achieving disentanglement and purification of modality-common and -specific knowledge.

Effectiveness of the CKA module. The introduction of CKA brings **3.1%/4.0%** on mAP/R1 performance improvement, which is further enhanced to **4.5%/5.5%** when utilizing MCP and MSP modules together. It implies that aligning the disentangled cross-modal knowledge in two independent inter- and intra-modality feature spaces can effectively preserve the discriminability of multiple knowledge in a controllable manner. Thus, the originally conflicting modality-common and -specific knowledge can coexist in the lifelong

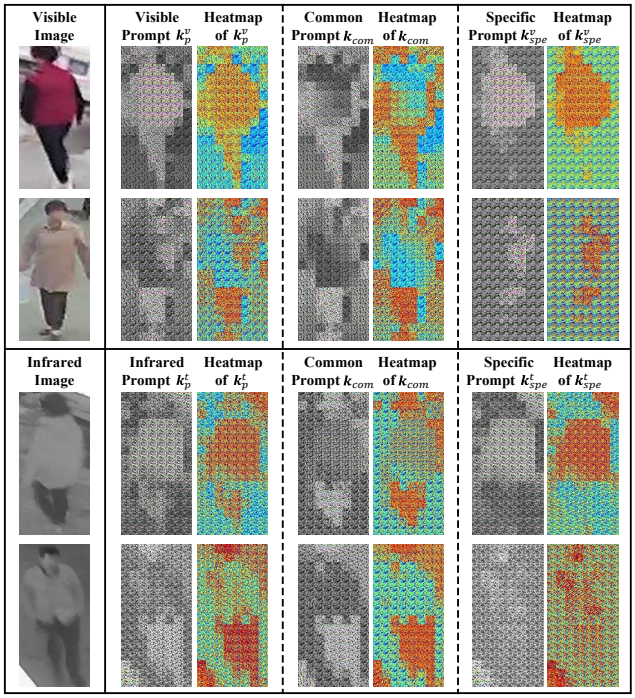


Figure 4: Visualization of the generated common prompt $\boldsymbol{k}_{com}$, specific prompt $\boldsymbol{k}_{spe}$, visible image prompt $\boldsymbol{k}^v$), and infrared image prompt $\boldsymbol{k}^v$.

process, alleviating the catastrophic forgetting derived from their mutual interference.

Hyperparameter Analysis. Our proposed CKDA introduces 4 hyper-parameters: α, β, μ in Eq. 15, and d in Eq. 3. Therefore, we evaluate their effects as shown in Figure 3. For α and β, our CKDA achieves the highest performance when both are set to 1. This is because an excessively large α prevents the model from separating the modality-common and -specific knowledge of the current data, while an excessively large β over-aligns the new and old knowledge, suppressing the discriminability of the new knowledge. Similarly, μ controls the alignment degree of modality-common and -specific knowledge, where the best performance is achieved at 0.5. This suggests a balance of equal importance between cross-modal knowledge, which further verifies that our CKDA can effectively separate the above knowledge in a balanced manner. Finally, the performance keeps improving until $d = 8$ and gradually decreases for larger values, which indicates the trade-off between the redundancy and difficulty of cross-modal knowledge, while our method balances them and selects the most appropriate number of channels. Therefore, according to the above analysis, the hyper-parameters α, β, μ, and d are set as 1, 1, 0.5 and 8, respectively.

Visualization

To intuitively demonstrate the effect of our method, we visualize the modality-common/-specific prompts with corresponding heatmaps as shown in Figure 4. It reveals that the common prompt tends to focus on the overall contour and body shape information of the person, which coexist in both modalities. In contrast, the specific prompt turns to focus on

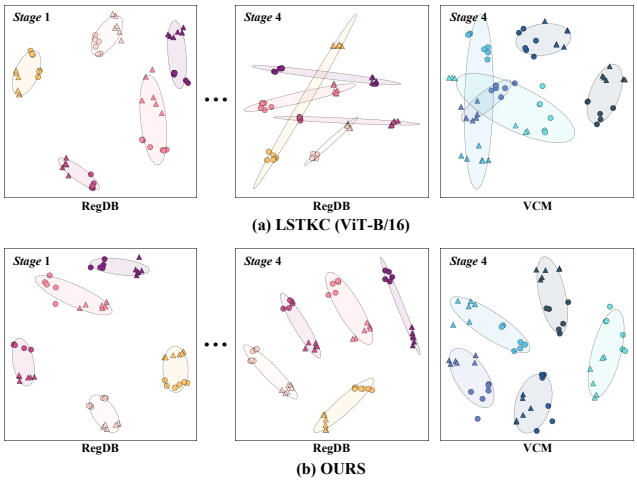


Figure 5: The t-SNE visualization of different training stages, where different colours represent different identities and ◯, △ represent features of infrared and visible modalities, respectively.

modality-specific knowledge that is only available in distinct modalities, *e.g.*, clothing color in visible images or heat-sensitive information in infrared images. As a result, the combined visible/infrared prompt can provide various perceptions in a complementary manner.

Additionally, Figure 5 visualizes the feature distribution extracted by our CKDA compared to LSTKC (Xu, Zou, and Zhou 2024) at different stages, including the old dataset RegDB and the new dataset VCM. Limited by mutual interference, LSTKC fail to alleviate the conflict between modality-common and -specific knowledge, thus leading to severe dual knowledge forgetting. While our CKDA not only promotes the purification of cross-modality knowledge by explicit disentangling, but also facilitates old knowledge retention by aligning it in a balanced way. Therefore, our CKDA achieves leadership in both new knowledge acquisition and old knowledge retention.

Conclusion

In this paper, we focus on a practical and challenging problem called Visible-Infrared Lifelong person ReIDentification (VI-LReID), which requires retrieving the same person based on sequentially collected cloth-consistent and cloth-changing data. To tackle the above issue, we propose a Modality-Common Prompting (MCP) module to balance cloth-relevant and cloth-irrelevant knowledge. Specifically, a Modality-Specific Prompting (MSP) module is designed to dynamically exploit differentiated new knowledge while eliminating the derived domain shift of old knowledge. Besides, a Cross-modality Knowledge Aligning (CKA) module is designed to alleviate the knowledge conflict between differentiated knowledge by aligning their distribution at different levels. Extensive experimental results illustrate that our CKDA outperforms existing methods in the challenging VI-LReID task.

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China (62525201, 62132001, 62432001) and Beijing Natural Science Foundation (L247006, L257005).

References

Ahmad, S.; Scarpellini, G.; Morerio, P.; and Del Bue, A. 2022. Event-driven re-id: A new benchmark and method towards privacy-preserving person re-identification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 459–468.

Ai, Z.; Cui, Z.; Peng, Y.; and Zhou, J. 2025a. UPP: Unified Point-Level Prompting for Robust Point Cloud Analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 27359–27368.

Ai, Z.; Liu, Z.; Lei, Y.; Cui, Z.; Zou, X.; and Zhou, J. 2025b. Gaprompt: Geometry-aware point cloud prompt for 3d vision model. *arXiv preprint arXiv:2505.04119*.

Chaudhry, A.; Dokania, P. K.; Ajanthan, T.; and Torr, P. H. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision*, 532–547.

Chen, H.; Lagadec, B.; and Bremond, F. 2022. Unsupervised lifelong person re-identification via contrastive rehearsal. *arXiv preprint arXiv:2203.06468*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Cui, Z.; Zhou, J.; and Peng, Y. 2024. Dma: Dual modality-aware alignment for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 19: 2696–2708.

Cui, Z.; Zhou, J.; Peng, Y.; Zhang, S.; and Wang, Y. 2023. Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. *IEEE transactions on circuits and systems for video technology*, 33(8): 4415–4428.

Cui, Z.; Zhou, J.; Wang, X.; Zhu, M.; and Peng, Y. 2024. Learning continual compatible representation for re-indexing free lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16614–16623.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3366–3385.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ge, W.; Du, J.; Wu, A.; Xian, Y.; Yan, K.; Huang, F.; and Zheng, W.-S. 2022. Lifelong person re-identification by pseudo task knowledge preservation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 688–696.

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

Hu, W.; Liu, B.; Zeng, H.; Hou, Y.; and Hu, H. 2022. Adversarial decoupling and modality-invariant representation learning for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8): 5095–5109.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, 448–456. pmlr.

Jiang, K.; Zhang, T.; Liu, X.; Qian, B.; Zhang, Y.; and Wu, F. 2022. Cross-modality transformer for visible-infrared person re-identification. In *European conference on computer vision*, 480–496. Springer.

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

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Liang, T.; Jin, Y.; Liu, W.; and Li, Y. 2023. Cross-modality transformer with modality mining for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 25: 8432–8444.

Lin, X.; Li, J.; Ma, Z.; Li, H.; Li, S.; Xu, K.; Lu, G.; and Zhang, D. 2022. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20973–20982.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lu, H.; Zou, X.; and Zhang, P. 2023. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 1835–1843.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 0–0.

Luo, Z.; Xiao, G.; Lew, M. S.; and Wu, S. 2025. Lifelong Visible-Infrared Person Re-Identification with Prompt Pool and Instance-level Prompt Generator. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 953–962.

Masana, M.; Liu, X.; Twardowski, B.; Menta, M.; Bagdanov, A. D.; and Van De Weijer, J. 2022. Class-incremental Learning: Survey and Performance Evaluation on Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5513–5533.

Moon, H.; and Phillips, P. J. 2001. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception*, 30(3): 303–321.

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

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

Pu, N.; Chen, W.; Liu, Y.; Bakker, E. M.; and Lew, M. S. 2021. Lifelong Person Re-identification via Adaptive Knowledge Accumulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7901–7910.

Pu, N.; Zhong, Z.; Sebe, N.; and Lew, M. S. 2023. A Memorizing and Generalizing Framework for Lifelong Person Re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.

Ren, K.; and Zhang, L. 2024. Implicit discriminative knowledge learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 393–402.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Sun, Z.; and Mu, Y. 2022. Patch-based Knowledge Distillation for Lifelong Person Re-Identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 696–707.

Sun, Z.; and Zhao, F. 2023. Counterfactual attention alignment for visible-infrared cross-modality person re-identification. *Pattern Recognition Letters*, 168: 79–85.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, 5380–5389.

Wu, G.; and Gong, S. 2021. Generalising without forgetting for lifelong person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2889–2897.

Wu, Q.; Dai, P.; Chen, J.; Lin, C.-W.; Wu, Y.; Huang, F.; Zhong, B.; and Ji, R. 2021. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4330–4339.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 374–382.

Xing, Y.; Xiao, G.; Lew, M. S.; and Wu, S. 2024. Lifelong visible-infrared person re-identification via a tri-token transformer with a query-key mechanism. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 988–997.

Xu, K.; Jiang, C.; Xiong, P.; Peng, Y.; and Zhou, J. 2025a. Dask: Distribution rehearsing via adaptive style kernel learning for exemplar-free lifelong person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8915–8923.

Xu, K.; Liu, Z.; Zou, X.; Peng, Y.; and Zhou, J. 2025b. Long Short-Term Knowledge Decomposition and Consolidation for Lifelong Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, K.; Zhang, H.; Li, Y.; Peng, Y.; and Zhou, J. 2024a. Mitigate catastrophic remembering via continual knowledge purification for noisy lifelong person re-identification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5790–5799.

Xu, K.; Zhuo, F.; Li, J.; Zou, X.; and Zhou, J. 2025c. Self-Reinforcing Prototype Evolution with Dual-Knowledge Cooperation for Semi-Supervised Lifelong Person Re-Identification. *arXiv preprint arXiv:2507.01884*.

Xu, K.; Zou, X.; Peng, Y.; and Zhou, J. 2024b. Distribution-aware Knowledge Prototyping for Non-exemplar Lifelong Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16604–16613.

Xu, K.; Zou, X.; and Zhou, J. 2024. LSTKC: Long Short-Term Knowledge Consolidation for Lifelong Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16202–16210.

Yao, Y.; Liu, Z.; Cui, Z.; Peng, Y.; and Zhou, J. 2025. Selective visual prompting in vision mamba. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22083–22091.

Ye, M.; Shen, J.; J. Crandall, D.; Shao, L.; and Luo, J. 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, 229–247. Springer.

Yu, C.; Shi, Y.; Liu, Z.; Gao, S.; and Wang, J. 2023a. Lifelong person re-identification via knowledge refreshing and consolidation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3295–3303.

Yu, H.; Cheng, X.; Peng, W.; Liu, W.; and Zhao, G. 2023b. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11185–11195.

Zhang, Q.; Lai, C.; Liu, J.; Huang, N.; and Han, J. 2022. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7349–7358.

Zhang, Y.; and Wang, H. 2023. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2153–2162.

Zhao, B.; Xiao, X.; Gan, G.; Zhang, B.; and Xia, S.-T. 2020. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13208–13217.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, 1116–1124.

Zhou, J.; Zhu, K.; Cui, Z.; Liu, Z.; Zou, X.; and Hua, G. 2025. State Space Prompting via Gathering and Spreading Spatio-Temporal Information for Video Understanding. *arXiv preprint arXiv:2510.12160*.

Zhu, X.; Xiao, G.; Lew, M. S.; and Wu, S. 2025. Lifelong visible–infrared person re-identification via replay samples domain-modality-mix reconstruction and cross-domain cognitive network. *Computer Vision and Image Understanding*, 254: 104328.