Lifelong visible–infrared person re-identification via replay samples domain-modality-mix reconstruction and cross-domain cognitive network

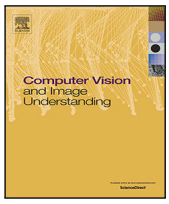
Zhu, X.; Xiao, G.; Lew, M.S.K.; Wu, S.

Computer Vision and Image Understanding 254 (2025) 104328

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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



Lifelong visible–infrared person re-identification via replay samples domain-modality-mix reconstruction and cross-domain cognitive network

Xianyu Zhu [a], Guoqiang Xiao [b], Michael S. Lew [c], Song Wu [b,d,*]

[a] Hanhong College, Southwest University, Chongqing, 400715, China
[b] College of Computer and Information Science, Southwest University, Chongqing, 400715, China
[c] LIACS Media Lab, Leiden University, Leiden, Netherlands
[d] Yibin Academy of Southwest University, Yibin, Sichuan, 644000, China

ARTICLE INFO

Keywords:
Lifelong learning
Visible–infrared person re-identification
Graph convolutional network

ABSTRACT

Adapting statically-trained models to the incessant influx of data streams poses a pivotal research challenge. Concurrently, visible and infrared person re-identification (VI-ReID) offers an all-day surveillance mode to advance intelligent surveillance and elevate public safety precautions. Hence, we are pioneering a more fine-grained exploration of the lifelong VI-ReID task at the camera level, aiming to imbue the learned models with the capabilities of lifelong learning and memory within the continuous data streams. This task confronts dual challenges of cross-modality and cross-domain variations. Thus, in this paper, we proposed a Domain-Modality-Mix (DMM) based replay samples reconstruction strategy and Cross-domain Cognitive Network (CDCN) to address those challenges. Firstly, we establish an effective and expandable baseline model based on residual neural networks. Secondly, capitalizing on the unexploited potential knowledge of a memory bank that archives diverse replay samples, we enhance the anti-forgetting ability of our model by the Domain-Modality-Mix strategy, which devising a cross-domain, cross-modal image-level replay sample reconstruction, effectively alleviating catastrophic forgetting induced by modality and domain variations. Finally, guided by the Chunking Theory in cognitive psychology, we designed a Cross-domain Cognitive Network, which incorporates a camera-aware, expandable graph convolutional cognitive network to facilitate adaptive learning of intra-modal consistencies and cross-modal similarities within continuous cross-domain data streams. Extensive experiments demonstrate that our proposed method has remarkable adaptability and robust resistance to forgetting and outperforms multiple state-of-the-art methods in comparative assessments of the performance of LVI-ReID. The source code of our designed method is at https://github.com/SWU-CS-MediaLab/DMM-CDCN.

1. Introduction

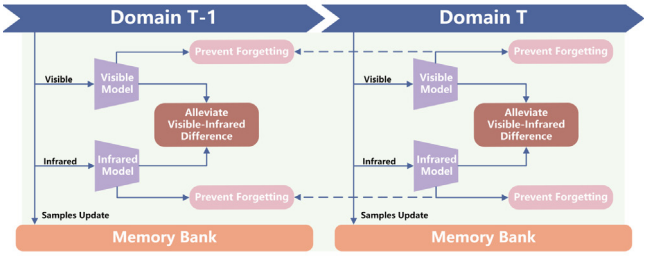
The Visible–Infrared Person Re-Identification (VI-ReID) task aims to identify the individual with the most matching identity from a database composed of images captured under both visible and infrared modalities. Unlike visible ReID tasks, which are confined to daytime applications, the VI-ReID task integrates both the visible and infrared modality, enabling all-day intelligent surveillance for security purposes, given its significant practical value. Many academic studies have approached enhancing the performance of VI-ReID from various perspectives (Wang et al., 2019b; Zhang and Wang, 2023; Li et al., 2020).

However, in practical applications of VI-ReID, there often exists an expansion and scalability requirement for the camera-based databases; meanwhile, these models trained solely on static datasets are not suitable in the real world where data continually pours in. As depicted in Figs. 1(c) and 1(d) under the joint learning scenario, if every new data domain necessitates retraining, it would incur substantial time and memory resource expenditures. Moreover, as depicted in Fig. 1(b), the model attains heightened generalization capability and adaptability through the continuous accrual of knowledge. Consequently, the VI-ReID model necessitates the capacity for lifelong learning to accommodate the ever-changing data domains.

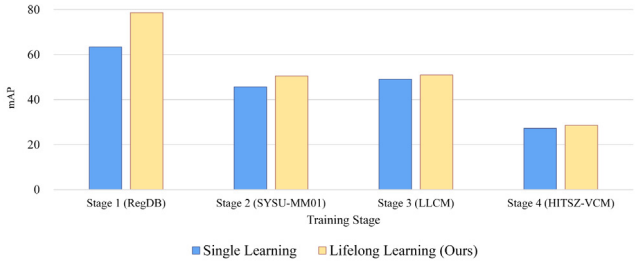
To achieve the aforementioned objective, as depicted in Fig. 1(a), TTQK (Xing et al., 2024) introduced the Lifelong Visible–Infrared Person Re-Identification (LVI-ReID). This task requires that the model can adapt to continuously arriving new data domains encompassing both visible and infrared modalities, while also maintaining its identification capabilities on previously encountered domains. We refer to these two key characteristics as domain adaptability and anti-forgetting ability. Addressing the requirements and characteristics of LVI-ReID, we face

* Corresponding author at: College of Computer and Information Science, Southwest University, Chongqing, 400715, China.
E-mail address: songwuswu@swu.edu.cn (S. Wu).

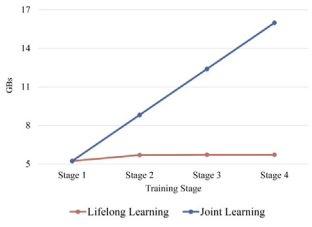
https://doi.org/10.1016/j.cviu.2025.104328
Received 13 May 2024; Received in revised form 28 January 2025; Accepted 24 February 2025
Available online 4 March 2025



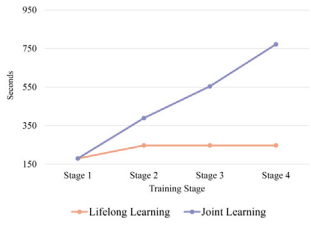
(a) Pipeline of the proposed LVI-ReID task. The model is required to continuously learn knowledge from incoming domains, alleviate cross-modal discrepancies, and prevent catastrophic forgetting.



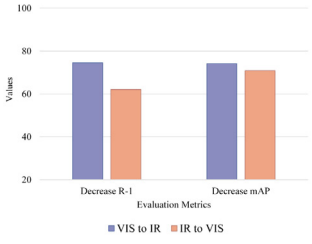
(b) Stage-wise mean Average Precision (mAP) under VIS (query) to IR (gallery) test mode for both Single-stage Learning and Multi-stage Lifelong Learning. Single Learning trains a model at a specific stage, whereas Lifelong Learning continuously updates it across diverse stages, accruing cross-domain knowledge. As a result, it surpasses Single Learning in terms of generalization and accuracy.



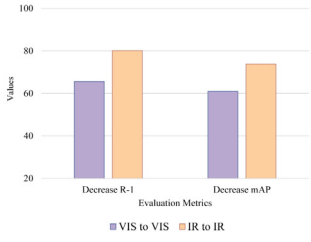
(c) Memory Cost for Each Stage (in GBs). A new dataset is added to the training at each stage, and Lifelong Learning consumes fewer resources compared to Joint Learning.



(d) Time Cost of Single Epoch (in Seconds). A new dataset is added to training at each stage. Lifelong learning takes less time than Joint Learning.



(e) Modal gaps in SYSU-MM01. VIS to IR means visible modality training with infrared modality testing, while IR to VIS involves infrared modality training with visible modality testing.



(f) Domain gaps from SYSU-MM01 to LLCM. We assessed domain gaps within Visible and infrared modalities across datasets, with more notable disparities in the infrared modality (IR to IR).

Fig. 1. Motivation of Our Work. To mitigate the substantial computational cost in terms of time and memory associated with retraining models at different stages (1(c) and 1(d)), we propose the LVI-ReID task (1(a)), endowing models with lifelong learning abilities that cumulatively accumulate knowledge for enhanced generalization and adaptability (1(b)). During the process of continual learning, we need to simultaneously manage the decline in model adaptability and catastrophic forgetting caused by modal gaps (1(e)) and domain gaps (1(f)).

multiple challenges: Firstly, we need to tackle the dual decline in model adaptability and anti-forgetting ability due to significant cross-modal discrepancies (as shown in Fig. 1(e)), which TTQK did not specifically address in its design. Secondly, as illustrated in Fig. 1(f), addressing the dual deterioration in adaptability and anti-forgetting ability caused by the domain semantic gap is another pressing challenge. Finally, TTQK does not leverage more fine-grained information, so how to fully utilize the potential of the learned model within environments characterized by dynamic expansion of camera networks and limited storage space is also an equally important challenge.

Prior studies have partly focused on realizing lifelong person re-identification under a single visible modality. For instance, AKA (Pu et al., 2021) implemented knowledge accumulation and manipulation by deploying the AKA network, while PTKP (Ge et al., 2022) employed pseudo-labeling techniques to mitigate the domain semantic gap between continuous learning data domains. However, due to the significant discrepancies across different data modalities and domains, these previous researches have not adequately addressed the multifaceted challenges inherent in LVI-ReID.

Thus, we propose an innovative LVI-ReID framework that aims to address these challenges from both the image and feature levels. At the image level, inspired by the principles of PCB (Sun et al., 2018) and CutMix (Yun et al., 2019), coupled with the inherent advantages of Memory Banks, we introduce the Domain-Modality-Mix (DMM) based replay samples reconstruction strategy. This strategy involves randomly mixing and reconstructing stored replay sample images across different partitioned regions with a certain probability, crossing domains and modalities. Operating the Domain-Modality-Mix strategy within a fixed storage space can amplify the diversity of replay samples, thereby effectively mitigating catastrophic forgetting caused by cross-domain and cross-modality semantic gaps.

Moreover, at the feature level, drawing inspiration from Chunking Theory (Gobet et al., 2001) in cognitive psychology and leveraging the robust feature aggregation properties of Graph Convolutional Networks (GCNs) (Jiang et al., 2019), we designed a novel Cross-domain Cognitive Network (CDCN) to promote the extraction of cross-modal invariant features. By implementing a camera-aware, dynamically expanding GCN framework, the cross-domain cognitive network harmoniously integrates visible and infrared data. This integration not only bridges the cross-modal disparity across disparate camera types but also inherently consolidates multi-domain information, leading to a substantial boost in the LVI-ReID performance of the learned model. Extensive experiments show the effectiveness of our proposed methodology. The main contributions of our proposed method can be summarized as follows:

- **Replay Samples Domain-Modality-Mix Reconstruction**: During the rehearsal phase in LVI-ReID, we innovatively designed a Domain-Modality-Mix (DMM) based replay sample reconstruction strategy. The DMM can generate diversified cross-domain and cross-modality samples as data augmentation, effectively serving as a countermeasure against catastrophic forgetting in LVI-ReID.
- **Cross-domain Cognitive Network**: To enhance the domain adaptability of the learned LVI-ReID model for the continuous cross-domain data streams, we designed a Cross-domain Cognitive Network (CDCN) framework, which is based on a camera-extended graph convolutional network to fuse cross-modal information and integrate cross-domain information at the feature level, having the advantages of real-world applications.
- **LVI-ReID Task Performance**: We tackle the LVI-ReID task at a more fine-grained level and explore a highly extensible baseline for the LVI-ReID task. Extensive experiments demonstrate that our proposed DMM strategy and CDCN framework have remarkable domain adaptability and robust resistance to forgetting and outperform multiple state-of-the-art methods in comparative assessments of the performance of LVI-ReID.

2. Related work

2.1. Visible–infrared person re-identification

The primary objective of the Visible–Infrared Person Re-identification (VI-ReID) task is to retrieve the most matching target individual from different modalities of images. Recent research can be mainly categorized into two categories: (1) The first category of methods aims to mitigate modality discrepancies by aligning or transforming images at the pixel level. Some methods employ Generative Adversarial Networks (GANs) to facilitate mutual conversion between visible and infrared modalities, thereby achieving alignment across modalities (Wang et al., 2019b; Choi et al., 2020; Wang et al., 2019a; Wu et al., 2024a; Liu et al., 2024). However, due to the scarcity of paired visible and infrared samples, GANs training inevitably introduces additional noise. To address this, some methods adopt channel-level data augmentation techniques (Ye et al., 2021; Hua et al., 2023) or utilize lightweight networks (Li et al., 2020; Wei et al., 2021; Zhang et al., 2021; Zhong et al., 2022; Kansal et al., 2020) to generate intermediate modality images, thereby reducing the significant differences between distinct modalities. However, these pixel-level methods mostly focus on a single data domain and do not take into account the mixture of multiple data domains. In contrast, our DMM fusion strategy has been specifically designed to handle data augmentation tasks that involve multiple data domains. (2) The second category of methods focuses on learning modality-invariant features to enable cross-modal retrieval, which is similar to cross-modal hashing (Wu et al., 2024b). For instance, Deep Zero-Padding Network (Wu et al., 2017a) automatically evolves modality-specific nodes . While some approaches (Ye et al., 2020; Wu et al., 2020; Lu et al., 2020) learn cross-modal invariant features by leveraging similarities across modalities or samples. MAUM (Liu et al., 2022) utilizes Memory-Augmented Unsupervised Metric Learning to strengthen cross-modal associations. Furthermore, DEEN (Zhang and Wang, 2023) achieves alignment between the distributions of visible and infrared modalities by learning rich feature representations through an embedding space enhancement network.

These VI-ReID models have been primarily researched and developed within the context of static datasets, with no explicit consideration given to their ability to learn over extended periods or across the lifetime dimension, thus indicating a certain gap between their current state and practical real-world applications where continuous learning is imperative.

2.2. Lifelong learning

Lifelong Learning, also known as continuous learning, envisions a model capable of effectively mastering the dual challenge of acquiring new knowledge while retaining previously learned information, which we denote as adaptability and anti-forgetting ability. The field's strategies can be broadly grouped into three categories: By employing an innovative replay sampling strategy, a subset of approaches (Rebuffi et al., 2017; Bang et al., 2021; Vitter, 1985; Wang et al., 2022; Zhu et al., 2020) select representative samples that aid the model in solidifying its grasp on previously learned knowledge. Other research efforts (Kong et al., 2023; Li and Hoiem, 2018; Kirkpatrick et al., 2016; Aljundi et al., 2018; Rannen et al., 2017) lean on regularization methods, expecting the teacher model to instruct the student model via knowledge distillation, thereby helping the student model to combat catastrophic forgetting effectively. Some methods (Serra et al., 2018; Yoon et al., 2017) adopt parameter isolation methods, they reserves the most valuable parameters for different tasks, thereby enhancing the accuracy of the model in performing earlier tasks. Additionally, addressing sample imbalance during Lifelong Learning is a concern for a part of approaches (Feng et al., 2023; Wu et al., 2019; Hou et al., 2019), ensuring fairness in the multi-step training.

Traditional Lifelong Learning tasks often focus their research on the closed dataset, where they struggle to perform well under the condition of continuously evolving data domains. In contrast, LVI-ReID task bridges the gap between research and practical applications by adapting to cross-domain and cross-modal datasets, thereby bringing the study of lifelong learning closer to real-world deployment.

2.3. Lifelong person re-identification

Lifelong Person Re-identification (LReID) is geared towards fulfilling the sustained demands of visual search systems. Notably, AKA (Pu et al., 2021) innovates by establishing a learnable knowledge graph, emulating human-like continuous learning through the accumulation and manipulation of knowledge. GwFReID (Wu and Gong, 2021) integrates a holistic learning objective, ensuring coherence across multiple objectives during training to enhance learning stability. PTKP (Ge et al., 2022) reframes LReID as an unsupervised domain adaptation challenge, achieving domain consistency by mapping features from novel tasks to those of prior tasks. Similarly, LUDA-ReID (Huang et al., 2022) leverages meta-learning and relational consistency to address catastrophic forgetting and assimilate new knowledge in continuous scenarios. KRKC (Yu et al., 2023) uniquely combines knowledge refreshing and rehearsal strategies, deploying a dynamic memory model alongside an adaptive working model to foster bidirectional knowledge exchange.

However, reliance on a single visible modality alone is insufficient for realizing around-the-clock intelligent surveillance. The inclusion of infrared modalities in LVI-ReID has made it possible to attain such functionality. Concurrently, previous studies with sample replay have overlooked the crucial role of diversity within the replayed samples in combating catastrophic forgetting, and they have not fully exploited the inherent advantages of Memory Banks to integrate cross-domain samples effectively.

2.4. Lifelong visible–infrared person re-identification

TTQK (Xing et al., 2024) first introduced the LVI-ReID task, and proposed a Tri-Token Transformer with a Query-Key mechanism to address the LVI-ReID task. However, TTQK focuses solely on solving the lifelong learning aspect and does not specifically address the modality differences in VI-ReID. Additionally, it lacks the capability to handle more fine-grained information, such as variations between different cameras.

2.5. Graph convolutional networks

In the field of person re-identification (ReID), Graph Convolutional Networks (GCNs) have seen extensive application, such as in supervised person ReID (Shen et al., 2018), unsupervised multi-source domain adaptation for person ReID (Bai et al., 2021), and lifelong person ReID (Pu et al., 2021).

However, there is a lack of research concerning GCN architectures that effectively reconcile the reduction of modal disparities and the concurrent accumulation of knowledge for LVI-ReID task.

3. Methodology

3.1. Problem definition and formulation

In Lifelong Visible–Infrared Person Re-Identification (LVI-ReID), the model needs to learn knowledge from a continual data stream containing T domains. Suppose there is a stream of datasets $D = \{D^i, C^i\}_{i=1}^{T}$, where $D^i = \{D^i_{train}, D^i_{test}\}$ contains the training and testing set and $C^i = \{C^{i,m}\}, m \in \{vis, ir\}$ involves the visible and infrared cameras set of the ith domain. Specifically, at each stage t, the tth training set $D^t_{train} = \{(x_i^{t,m}, y_i^t, c_i^{t,m})\}_{i=1}^{|D^t_{train}|}, m \in \{vis, ir\}$ contains the images

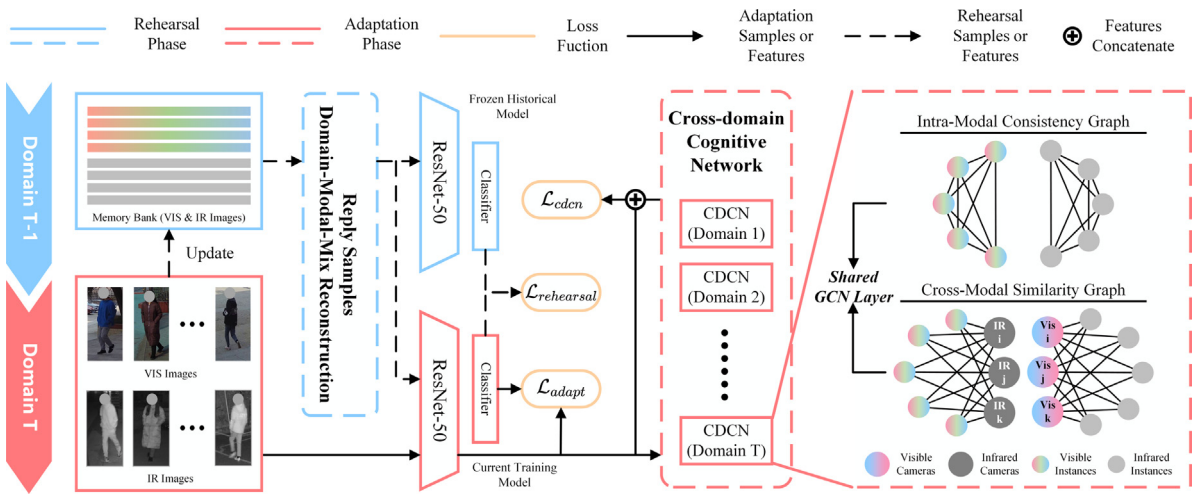


Fig. 2. The overview of our proposed methods. In the LVI-ReID task, a training stage comprises two phases, an adaptation phase (in red) and a rehearsal phase (in blue). During the adaptation phase, a mini-batch sampled from current domain is processed initially through the backbone network. The resultant features are subsequently fed into our camera-aware Cross-domain Cognitive Network, tasked with accumulating knowledge, ensuring intra-modal consistency, and mining cross-modal similarities. The adaptability of our model is refined through a collaborative optimization process steered by the loss functions $\mathcal{L}_{adapt}$ and $\mathcal{L}_{cdcn}$. Transitioning into the Rehearsal Phase, a mini-batch of images retrieved from the memory bank is reconstructed via our Replay Samples Domain-Modal-Mix Reconstruction. This operation facilitates cross-modal and cross-domain integration. The resultant rehearsal instances are mapped into confidence scores independently by both a temporally frozen model, retained from the previous stage, and the current training model. Knowledge preservation is enforced through $\mathcal{L}_{rehearsal}$, which enables a distilled knowledge transfer from the frozen to the training model, mitigating the issue of catastrophic forgetting. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$x_i^{t,m}$ from either the visible or infrared modality in the training set of the tth domain, y_i^t is the corresponding identity label and $c_i^{t,m}$ is the corresponding camera label belonging to $C^{t,m}$. As the setting in previous lifelong person re-identification works, D_{train}^t is available only at the tth stage, and we build up an exemplar memory M^t to store a limited number of samplers from each stage for rehearsal. For inference, we will evaluate the adaptability and anti-forgetting ability of the trained model using all testing sets from all seen domains $\{D_{test}^i\}_{i=1}^t$ in the LVI-ReID task.

3.2. Method overview

As depicted in Fig. 2 , our method consists of two main processes when training the tth domain at the tth stage: the adaptation phase (red) and the rehearsal phase (blue). During the adaptation phase, a batch of samples from the training set D_{train}^t will first be mapped into high-dimensional features by the Backbone network, followed by projection onto a confidence interval by the Classifier, and subjected to constraints imposed by an adaptation loss function $\mathcal{L}_{adapt}$. Subsequently, we propose leveraging a Cross-domain Cognitive Network (CDCN) to facilitate model learning of intra-modal consistency, meanwhile constructing distinct cross-modal similarity graphs for each data domain based on their respective sets of cameras C^t, thereby mitigating the impact of modality semantic gaps and enhancing the capability of the learned model to recognize cross-modal similarities. Moreover, a shared GCN layer is employed to alleviate the domain semantic gap and reinforce the cross-domain generalization ability and adaptability of the learned model. During the rehearsal phase, we capitalize on the inherent strengths of a memory bank and propose the Domain-Modality-Mixing (DMM) based replay samples strategy, which performs cross-domain and cross-modal sample reconstruction of replay samples. As shown in Fig. 3, the strategy of DMM commences by vertically partitioning a three-channel replay sample into several blocks, where, for each block, it first opts for one of three channel mixing ways: (1) Single-channel Mixing, (2) Dual-channel Mixing, (3) All-channel Mixing. Following the selection of channel mixing strategies, DMM proceeds with a certain probability to implement any of the following four blending approaches: (1) Intra-Domain Same-Modality Mixing Samples, (2) Intra-Domain Cross-Modality Mixing Samples, (3) Cross-Domain Same-Modality Mixing Samples, (4) Cross-Domain Cross-Modality Mixing Samples. Through channel-level Domain-Modality-Mixing, the diversity of replay samples is significantly augmented, thereby profoundly mitigating catastrophic forgetting issues induced by both domain transfer and cross-modal disparity. As a result, this approach substantially enhances the model's anti-forgetting ability. Then, the reconstructed samples are mapped onto a confidence interval by the frozen model from the previous stage and the current training model. Furthermore, knowledge distillation is performed using the rehearsal loss function $\mathcal{L}_{rehearsal}$, encouraging the model to combat catastrophic forgetting effectively.

3.3. Baseline approach

We establish a baseline LVI-ReID approach by incorporating knowledge distillation (Li and Hoiem, 2018) and utilizing a replay-based strategy underpinned by iCaRL (Rebuffi et al., 2017) methods. The baseline model consists of a feature extractor $H(\cdot;\theta)$ with parameters θ and an incremental classifier $G(\cdot;\phi)$ with parameters ϕ whose output dimension can expand as the number of total seen identities. Similarly to previous VI-ReID works, we add a Batch Normalization layer before the classifier, denoted as $BN(\cdot)$. So, the whole network defined as $F(\cdot;\theta;\phi) = G(BN(H(\cdot;\theta));\phi)$ is the mapping from the input space to confidence scores. When training new samples, we use cross-entropy loss $\mathcal{L}_{id}$ and triplet loss $\mathcal{L}_{tri}$ to conduct the adaptation loss function $\mathcal{L}_{adapt}$ and optimize the adaptability of the model:

$$\mathcal{L}_{id} = -\frac{1}{N_b^n}\sum_{i=1}^{N_b^n}\sum_{j=1}^{N^P}\hat{y}_{ij}\log\sigma(F^n(y_{ij}|x_i;\theta;\phi)) \tag{1}$$

where N_b^n represents the mini-batch size during the adaptation phase, while N^P represents the total number of classes. The one-hot encoded label for sample i is given by y_{ij}. Here, σ is softmax function, and $F^n(\hat{y}_{ij}|x_i;\theta;\phi)$ represents the training model's predicted probability that sample x_i belongs to class j.

$$\mathcal{L}_{tri} = \frac{1}{N_b^n}\sum_{i=1}^{N_b^n}\Big[max\Big(d\big(H^n(a_i),H^n(p_i)\big) - d\big(H^n(a_i),H^n(r_i)\big)+m,0\Big)\Big] \tag{2}$$

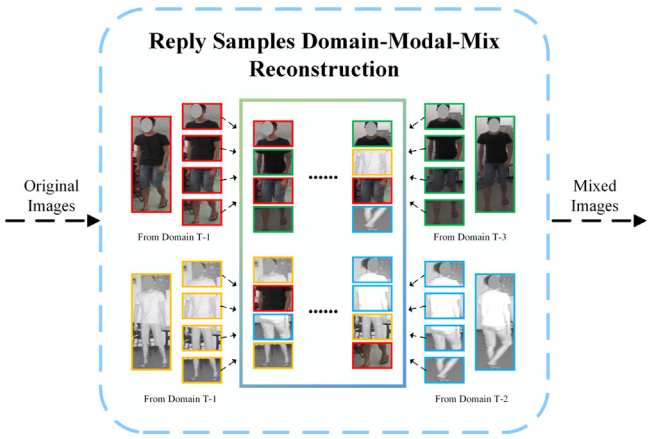


Fig. 3. The Replay Samples Domain-Modal-Mix Strategy. The original replay images from different domains and modalities will be mixed according to certain rules to enhance the diversity of replay samples, thereby reducing the performance degradation of the model during training caused by domain gaps and modal gaps.

where N_b^n is the mini-batch size during the adaptation phase, while d represents the Euclidean distance. $H^n(\cdot)$ is the training feature extractor, a_i stands for the anchor sample, p denotes a positive sample of the anchor, r denotes a negative sample of the anchor and m is a constant. Then, we can obtain the optimization adaptation loss function $\mathcal{L}_{adapt}$ in the adaptation phase of model learning:

$$\mathcal{L}_{adapt} = \mathcal{L}_{id} + \mathcal{L}_{tri} \tag{3}$$

During the rehearsal phase, initially, we follow the replay sampling strategy from iCaRL (Rebuffi et al., 2017), updating the memory bank after the completion of training at each incremental stage. Furthermore, we apply knowledge distillation (Li and Hoiem, 2018) to the replay samples and conduct the rehearsal loss function $\mathcal{L}_{rehearsal}$, aiming to bolster the baseline model's resilience against forgetting:

$$\mathcal{L}_{rehearsal} = \frac{T^2}{N_b^o} \sum_{i=1}^{N_b^o} KL(F^n(x_i) || F^o(x_i)) \tag{4}$$

where N_b^o is the size of a rehearsal mini-batch, and T represents the temperature coefficient. $KL(\cdot \parallel \cdot)$ denotes the Kullback–Leibler divergence, $F^n(\cdot)$ is the training network's output, and $F^o(\cdot)$ is the frozen old network's output with gradient updates stopped. Ultimately, we can get the total optimization object for the Baseline:

$$\mathcal{L}_{base} = \mathcal{L}_{adapt} + \mathcal{L}_{rehearsal} \tag{5}$$

3.4. Replay samples domain-modality-mix reconstruction

In this section, we will introduce the details of the proposed Replay Samples Domain-Modality-Mix Reconstruction. Given that VI-ReID constitutes a fine-grained problem, overfitting the old data domain is prone to occur when relying solely on a small, repetitive set of replayed samples, exacerbating catastrophic forgetting. By contrast, a diverse set of replay samples has the potential to cross a broader range of feature spaces, thereby enabling the model to revisit critical features and patterns across various tasks while continuously learning new ones, thereby enhancing its ability to retain and discriminate among multiple categories learned in the past. Moreover, previous LReID tasks overlooked the potential knowledge in the memory bank containing replay samples from multiple domains. The implementation of cross-domain sample mixing during sample replay not only serves to mitigate further the detrimental effects of domain semantic gaps but also, by leveraging frozen legacy models and knowledge distillation techniques, eliminates concerns regarding the misguidance of the model due to inappropriate labeling of mixed images, ensuring that the model learns accurately even from such composite data.

To address the problems mentioned above in LVI-ReID, we draw inspiration from the methods of PCB (Sun et al., 2018), CutMix (Yun et al., 2019), and CA (Ye et al., 2021). Under the constraint of maintaining the same memory size, we propose the Domain-Modality-Mix (DMM) based replay samples reconstruction strategy to achieve the cross-domain sample mix and improve the diversity of replay samples, thereby effectively mitigating the issue of catastrophic forgetting in the learned model. Our DMM strategy involves two main selection strategies: Channel Mixing Strategies and Sample Mixing Strategies. The channel mixing operation can be denoted as $CH(\cdot,\cdot;ch;ct)$, where the first parameter refers to the primary image block, while the second parameter signifies the alternative sample block, $ch \in \{1,2,3\}$ represents different channel mixing tactics and $ct \in \{1,2,3\}$ denotes the different channel types selected. For any two sample blocks with the same vertical position $x_{bk,0} = \{x_{bk,0}^R, x_{bk,0}^G, x_{bk,0}^B\}$ and $x_{bk,1} = \{x_{bk,1}^R, x_{bk,1}^G, x_{bk,1}^B\}$, each consisting of three RGB channels (with identical RGB channel values for infrared samples), the channel mixing operation can be mathematically expressed as follows:

$$CH(x_{bk,0}, x_{bk,1}; 1; ct) = \begin{cases} \{x_{bk,1}^R,\ x_{bk,0}^G,\ x_{bk,0}^B\}, & ct = 1 \\ \{x_{bk,0}^R,\ x_{bk,1}^G,\ x_{bk,0}^B\}, & ct = 2 \\ \{x_{bk,0}^R,\ x_{bk,0}^G,\ x_{bk,1}^B\}, & ct = 3 \end{cases} \tag{6}$$

$$CH(x_{bk,0}, x_{bk,1}; 2; ct) = \begin{cases} \{x_{bk,1}^R,\ x_{bk,1}^G,\ x_{bk,0}^B\}, & ct = 1 \\ \{x_{bk,0}^R,\ x_{bk,1}^G,\ x_{bk,1}^B\}, & ct = 2 \\ \{x_{bk,1}^R,\ x_{bk,0}^G,\ x_{bk,1}^B\}, & ct = 3 \end{cases} \tag{7}$$

$$CH(x_{bk,0}, x_{bk,1}; 3; ct) = x_{bk,1},\ ct \in \{1,2,3\} \tag{8}$$

where $ch = 1$ represents single-channel mixing across the three RGB channels, $ch = 2$ signifies the random selection of two out of the three RGB channels for mixing, and $ch = 3$ indicates a complete replacement using $x_{bk,1}$.

After selecting channel mixing strategies, DMM will adopt sample mixing strategies. The sample mixing strategy can be represented as $SDM(\cdot,\{\};sd;sm;st;ch;ct)$, where the initial argument is the primary sample block, followed by a set of alternative sample blocks. Here, $sd \in \{0,1\}$ indicates whether cross-domain mixing is to be performed, while $sm \in \{0,1\}$ signifies if cross-modal mixing will perform. The variable st denotes the number of domains whose samples are stored in the memory bank M^t, while ch and ct are channel mixing ways and channel types, respectively. Assume that when $st < 2$, for any given sample block $x_{bk,0}$, there exists a set of sample blocks comprising the same-domain and same-modality block $x_{bk,1}^{ss}$ and the same-domain and cross-modality block $x_{bk,2}^{sc}$, both aligned vertically. Conversely, when $st \geqslant 2$, for any sample block $x_{bk,0}$, the associated set encompasses the same-domain and same-modality block $x_{bk,1}^{ss}$, the same-domain and cross-modality block $x_{bk,2}^{sc}$, the cross-domain and same-modality block $x_{bk,3}^{cs}$, and the cross-domain and cross-modality block $x_{bk,4}^{cc}$, all sharing the same vertical alignment (it is noteworthy that the same-domain and same-modality block $x_{bk,1}^{ss}$ could be $x_{bk,0}$ itself, implying no actual mixing in this instance). The sample mixing strategy can then be mathematically formalized as follows:

$$SDM(x_{bk,0}, \{x_{bk,1}^{ss}, x_{bk,2}^{sc}\}; sd; sm; st < 2; ch; ct) = \begin{cases} CH(x_{bk,0}, x_{bk,1}^{ss}; ch; ct), & sd = 0, sm = 0 \\ CH(x_{bk,0}, x_{bk,2}^{sc}; ch; ct), & sd = 0, sm = 1 \end{cases} \tag{9}$$

$$SDM(x_{bk,0}, \{x_{bk,1}^{ss}, x_{bk,2}^{sc}, x_{bk,3}^{cs}, x_{bk,4}^{cc}\}; sd; sm; st \geqslant 2; ch; ct) = \begin{cases} CH(x_{bk,0}, x_{bk,1}^{ss}; ch; ct), & sd = 0, sm = 0 \\ CH(x_{bk,0}, x_{bk,2}^{sc}; ch; ct), & sd = 0, sm = 1 \\ CH(x_{bk,0}, x_{bk,3}^{cs}; ch; ct), & sd = 1, sm = 0 \\ CH(x_{bk,0}, x_{bk,4}^{cc}; ch; ct), & sd = 1, sm = 1 \end{cases} \tag{10}$$

During the rehearsal phase, we select multiple replay samples from the memory bank M^t to form a rehearsal mini-batch equivalent to an adaptation mini-batch. For each replay sample $x_r^{t,m}$ belonging to tth domain and m modality. DMM first divide it into k blocks vertically $x_r^{t,m} = \{x_{r,j}^{t,m}\}_{j=1}^{k}$. Subsequently, DMM proceeds to perform domain mix with probability p_{cd} and modality mix with probability p_{cm} for each block $x_{r,j}^{t,m}$. The algorithmic process of DMM is illustrated in Algorithm 1.

Algorithm 1: Replay Samples Domain-Modality-Mix Reconstruction

Input : Replay Sample $x_r^{t,m}$, Memory Bank M^t,
Number of Domains st in Memory Bank ,
Number of Blocks per Sample k,
Modality Mix Probability p_{cm},
Domain Mix Probability p_{cd},
Mixing Operation $SDM(\cdot, \{\}; sd; sm; st; ch; ct)$
Output: Reconstruction Sample x_r^{rsdmm}

```
// Divide into k blocks
```
$\{x_{r,j}^{t,m}\}_{j=1}^{k} \leftarrow x_r^{t,m}$;
for $j = 1$ *to* k **do**
  $ch \leftarrow Random\_Choice\ (1, 2, 3)$;
  $ct \leftarrow Random\_Choice\ (1, 2, 3)$;
  $sd \leftarrow Random\_Choice\_with\_probability\ (0, 1; p_{cd})$;
  $sm \leftarrow Random\_Choice\_with\_Probability\ (0, 1; p_{cm})$;
  if $st \geqslant 2$ **then**
    $Atset \leftarrow \{x_{bk,1}^{ss}, x_{bk,2}^{sc}, x_{bk,3}^{cs}, x_{bk,4}^{cc}\} \in M^t$;
  else
    $Atset \leftarrow \{x_{bk,1}^{ss}, x_{bk,2}^{sc}\} \in M^t$;
  end
  $x_{r,j}^{rsdmm} \leftarrow SDM(x_{r,j}^{t,m}, Atset; sd; sm; st; ch; ct)$;
end
```
// Combine blocks
```
$x_r^{rsdmm} \leftarrow \{x_{r,j}^{rsdmm}\}_{j=1}^{k}$;
return x_r^{rsdmm}

3.5. Cross-domain cognitive network

In this section, we will introduce the details of the proposed Cross-domain Cognitive Network (CDCN). In the LVI-ReID task, alleviating the performance degradation caused by the modality semantic gap poses a significant challenge. Since Graph Convolutional Networks (Jiang et al., 2019) have powerful information aggregation and transmission ability, they are employed to enhance our model's cross-modal and cross-domain cognition ability. Inspired by AKA (Pu et al., 2021), we first construct consistency graphs within visible and infrared modalities to encourage the model to learn intra-modal consistency. Furthermore, given the abundance of information encapsulated within cameras in VI-ReID tasks, we adopt the Chunking Theory (Gobet et al., 2001) from cognitive psychology to conduct the cognitive network. We aggregate information from within individual cameras into cognitive nodes, thereby packaging camera-specific details, and establish cross-modal similarity graphs via these cognitive nodes. This strategy facilitates cross-modal information transfer and promotes learning of cross-modal invariance while accumulating knowledge. Additionally, leveraging domain-specific cognitive nodes, we construct tailored cross-modal similarity graphs for each domain, encouraging the model to address intra-domain cross-modal cognitive disparities specifically. However, focusing solely on intra-domain cross-modal cognition is insufficient, and the performance decrement induced by the domain semantic gap must also be addressed. To this end, we employ a shared GCN Layer to facilitate the integration of cross-domain information.

Finally, we design a camera-aware cross-domain cognitive network (equipping with a batch-norm layer $bn()$, a classifier $g(\cdot; \phi^f)$, and a shared GCN Layer) for each stage. The proposed CDCN framework contains three processes: (1) knowledge accumulation, (2) cognitive graph construct, and (3) knowledge transfer.

3.5.1. Knowledge accumulation

For our cognitive nodes' knowledge acquiring, we use movement average to accumulate knowledge from samples belonging to the same camera in each mini-batch. In the course of either the adaptation phase or the inference of stage t, we gain access to the camera label set $C^{t,m} = \{c_j^{t,m}\}_{j=1}^{|C^{t,m}|}, m \in \{vis, ir\}$ belonging to the tth domain. Specifically, during the adaptation phase, for each individual camera $c_j^{t,m}$, we systematically generate and permanently store a dedicated cognitive node $n_j^{t,m}$ equipped with a fully-connected layer denoted as $FC_j^{t,m}$. Moreover, when considering a set of sample instances $\{x_i\}_{i=1}^{N_c}$ collectively associated with camera $c_j^{t,m}$, the value $v_j^{t,m}$ of the corresponding cognitive node $n_j^{t,m}$ undergoes an update governed by the momentum-averaging mechanism:

$$v_j^{t,m} = (1 - mtm) \times \sum_{i=1}^{N_c} w_i H(x_i; \theta) + mtm \times v_j^{t,m} \tag{11}$$

where $H(x_i; \theta)$ symbolizes the feature extractor. The weight assigned to the extracted feature $H(x_i; \theta)$ is denoted by w_i, which is derived through the fully connected layer $FC_j^{t,m}$ and mtm is the moving coefficient which conventionally set to a value of 0.9. In this way, we can constantly accumulate knowledge and get a group of updated values $V_t^S = \{v_j^{t,m}\}_{j=1}^{|C^{t,m}|}, m \in \{vis, ir\}$ consisting of all cognitive nodes of the tth domain. It should be noted that the cognitive nodes themselves do not learn cross-modality similarity directly, they are the anchors for the GCN layer to learn cross-modality similarity, we design cognitive nodes to accumulate knowledge in the camera level and help GCN layer to optimize parameters and learn cross-modality similarity.

3.5.2. Cognitive graph construct

The primary goal of the CDCN is to learn the intra-modal similarities and cross-modal consistencies. To achieve this, we employ a dual-graph architecture $\mathcal{G} = (A, V)$ consisting of two unique components: an instance-based Intra-Modal Consistency Graph (IMCG), designed to integrate and harmonize knowledge within the same modality, and a domain-specific Cross-Modal Similarity Graph (CMSG), designed to reconcile disparities between visible and infrared modalities within the same domain.

Intra-Modal Consistency Graph. Given a mini-batch from the adaptation or inference phase, the N_m features extracted from either the visible or infrared modality are defined as $V^{C,m} = H(X^m; \theta)$, where $X^m = \{x_i^m\}_{i=1}^{N_m},\ m \in \{vis, ir\}$. Inspired by AKA (Pu et al., 2021), we employ the IMCG $\mathcal{G}^C(A^C, V^C)$ to discover the intra-modal relationship among instances across visible and infrared modalities, respectively. More specifically, we define the edge set of each modality as $A^{C,m}$, which is fully connected, while $V^{C,m}$ represents the feature values extracted from the backbone consisting of instances belonging to m modal. We regard the contributions from instance i_m to instance j_m as equal in the intra-modal, so we set the equal edge weights between different instances in the same modality:

$$\begin{aligned} &A_{i_m j_m}^{C,m} = 1.0,\ i_m j_m \in [1, N_m],\ m \in \{vis, ir\} \\ &A^C = \begin{bmatrix} A^{C,vis} & 0 \\ 0 & A^{C,ir} \end{bmatrix},\quad V^C = \begin{bmatrix} V^{C,vis} \\ V^{C,ir} \end{bmatrix} \end{aligned} \tag{12}$$

where $A^{C,m} \in \mathbb{R}^{N_m \times N_m}$, and $A^C \in \mathbb{R}^{2N_m \times 2N_m}$ gives the adjacency matrix of IMCG $\mathcal{G}^C$, while $V^C \in \mathbb{R}^{2N_m \times d}$ is concatenated by d dimension feature values $V^{C,m} \in \mathbb{R}^{N_m \times d}$ from both visible and infrared modalities.

Cross-modal Similarity Graph. In order to address disparities between visible and infrared modalities within the same domain, we adopt the Cross-modal Similarity Graph (CMSG) $\mathcal{G}^K(A^K, V^K)$ to learn the cross-modal invariance based on visible and infrared cognitive nodes. Specifically, in the tth domain, for a mini-batch of size N_{vis} sampled from visible modality, each instance is connected to every infrared cognitive node of the tth domain:

$$A_{i_{vis} k_{ir}}^{K,vis} = 1.0,\ i_{vis} \in [1, N_{vis}],\ k_{ir} \in [1, |C^{t,ir}|] \tag{13}$$

where i_{vis} is the index of a single instance in a visible mini-batch, and k_{ir} is the index of an infrared cognitive node in the domain, with $|C^{t,ir}|$ representing the total number of infrared cognitive nodes in the tth domain. It is similar to a mini-batch of size N_{ir} sampled from the infrared modality:

$$A^{K,ir}_{i_{ir},k_{vis}} = 1.0,\ i_{ir} \in [1, N_{ir}],\ k_{vis} \in [|C^{t,ir}|, |C^{t,ir}| + |C^{t,vis}|] \tag{14}$$

where k_{vis} is the index of a visible cognitive node in the domain and $|C^{t,vis}|$ represents the total number of visible cognitive nodes in the tth domain. Subsequently, we can get the adjacency matrix A^K of CMSG $\mathcal{G}^K$:

$$A^K = \begin{bmatrix} 0 & A^{K,vis} \\ A^{K,ir} & A^{K,eye} \end{bmatrix},\quad V^K = V^S_t \tag{15}$$

where $A^{K,vis} \in \mathbb{R}^{2N_m \times |C^t|}$, $A^{K,ir} \in \mathbb{R}^{|C^t| \times 2N_m}$ gives the relationship between instances and cross-modal cognitive nodes, $V^S_t \in \mathbb{R}^{|C^t| \times d}$ represents all values of cognitive nodes in the camera set C^t of tth domain, and $A^{K,eye}$ is an identity matrix meaning that every cognitive node is directly connected to itself.

3.5.3. Knowledge transfer

Aiming to stimulate the model's capacity to capture both intra-modal consistency and cross-modal similarity, we deploy a shared GCN layer to instantiate an environment where intra-modal instance nodes engage in knowledge sharing and interact with cross-modal cognitive nodes, all within a context of cross-domain knowledge integration.

Cognitive Graph. By merging A^C and A^K, V^C and V^K, we acquire the edge set A and value vector V of joint graph $\mathcal{G}$:

$$A = \begin{bmatrix} A^C & A^{K,vis} \\ A^{K,ir} & A^{K,eye} \end{bmatrix},\quad V = \begin{bmatrix} V^C \\ V^K \end{bmatrix} \tag{16}$$

where $A \in \mathbb{R}^{(2N_m+|C^t|)\times(2N_m+|C^t|)}$ indicates the correlation of intra-modal instances, and the relationship between instances and cross-modal cognitive nodes, with $V \in \mathbb{R}^{(2N_m+|C^t|)\times d}$ representing the value vector.

Knowledge Transfer. After constructing the joint graph $\mathcal{G}$, we transfer the intra-modal and cross-modal knowledge via the GCN, which is formulated as:

$$\hat{V} = \delta(A(VW)) \tag{17}$$

where $\hat{V}$ is the vertex embedding after one-layer graph convolutional operation, W is a learnable weight matrix of the GCN layer, and δ is a non-linear function, e.g., LeakyReLu (Xu et al., 2015). We employed only a single layer to accomplish knowledge transmission both within and across modalities, yet it is feasible to seamlessly superimpose multiple layers of GCN. After completing the knowledge transmission, we obtain the knowledge representation vector $\bar{V}$ from top $2N_m$ rows of $\hat{V}$, denoted as $\bar{V} = \{\hat{V}_i | i \in [1, 2N_m]\}$.

3.5.4. The optimization of CDCN

Utilizing residual connections, we obtain the final feature vectors V^F of CDCN outputs, employing the knowledge representation vector $\bar{V}$ and feature values V^C of instances for the process. Subsequently, we input the feature vectors V^{OF} sequentially into the CDCN's shared batch normalization layer $bn()$ and shared classifier $g(\cdot;\phi^c)$ to obtain confidence scores, and optimize CDCN using cross-entropy loss:

$$\mathcal{L}_{cdcn} = -\frac{1}{2N_m}\sum_{i=1}^{2N_m}\sum_{j=1}^{N^P} \hat{y}_{ij} \log \sigma(g(bn(V_i^F);\phi^c)) \tag{18}$$

where $2N_m$ is the size of mini-batch, N^P is the total number of classes, $g(bn(V_i^{OF});\phi^f)$ is the confidence scores, and $\sigma(\cdot)$ is the softmax function.

3.5.5. The inference of CDCN

This section details the distinctions between the training and inference phases of the CDCN framework. The inference phase is distinctive as it omits the accumulation of new knowledge and optimization procedures, and it primarily focuses on utilizing the accrued knowledge for predictions. By the way, during the rehearsal phase, CDCN does not participate actively due to the risk of incorporating misleading information from mixed samples, which could potentially corrupt its learned knowledge.

Note that, during the inference phase, as we deal with only a single modality for both the query and gallery sets, the edge set for a mini-batch of size N_e is represented as $A^C_e = A^{C,m} \in \mathbb{R}^{N_e \times N_e}$, the corresponding value vector as $V^C_e = V^{C,m} \in \mathbb{R}^{N_e \times d}$, while the edge set A^K_e, and values vector V^K_e of CSG are similar to adaptation phase. Ultimately, we can get the edge set and values vector of $\mathcal{G}_e(A_e, V_e)$:

$$A_e = \begin{bmatrix} A^{C,m}_e & A^{K,vis}_e \\ A^{K,ir}_e & A^{K,eye}_e \end{bmatrix},\quad V_e = \begin{bmatrix} V^{C,m}_e \\ V^K_e \end{bmatrix} \tag{19}$$

where $A_e \in \mathbb{R}^{(N_e+|C^t|)\times(N_e+|C^t|)}$, $V_e \in \mathbb{R}^{(N_e+|C^t|)\times d}$ and $m \in \{vis, ir\}$ represents the instances belonging to different modality from either the query set or the gallery set under different test mode.

3.5.6. Multi-loss optimization

It is worth noting that we utilize a single shared GCN layer throughout all training stages to learn and accumulate knowledge from different domains. For each domain, new cognitive nodes are created for each camera to memorize domain-specific knowledge at the camera level. During inference, these domain-specific cognitive nodes and the shared GCN layer work together to leverage the learned cross-domain knowledge. Finally, the optimization function of our model can be described as follows:

$$\mathcal{L} = \mathcal{L}_{adapt} + \mathcal{L}_{rehearsal} + \mathcal{L}_{cdcn} \tag{20}$$

4. Experiments

4.1. New benchmark for LVI-ReID

To evaluate the adaptability and resistance to forgetting the proposed model, we devise a continual learning data stream scheme **RegDB→SYSU→LLCM→VCM** based on four visible–infrared person re-identification datasets:

RegDB (Nguyen et al., 2017) comprises 412 unique individual identities, each including ten pairs of visible and infrared modality images. Following the practice in prior works, we randomly select 206 identities for training and reserve the remaining identities for testing.

SYSU-MM01 (Wu et al., 2017b) includes 491 distinct identities collected by four visible and two infrared cameras. The training set consists of 395 identities with 22258 visible images and 11909 infrared images. In the testing phase, this dataset provides two different search modes: Indoor Search and All-Search. Notably, we will exclusively employ the All-Search mode, wherein all images captured by visible cameras are incorporated into the gallery set.

LLCM (Zhang and Wang, 2023) is a low-light cross-modal dataset encompassing 1064 objects with unique identity labels. The image data was continuously collected over more than 100 days from January to April by nine RGB cameras and nine infrared cameras, spanning various weather conditions and clothing styles. Of these, 713 identities are allocated to the training set, while the remaining 351 identities form the test set (the number of query samples is 8680).

HITSZ-VCM (Lin et al., 2022) is a video-based visible–infrared person re-identification dataset that captures behavioral trajectories of 927 identities using six RGB cameras and six infrared cameras. Each identity sequence is constructed as a tracklet consisting of 24 consecutive frames, resulting in a total of 11785 visible tracklets and

Table 1
Adaptation performance under VIS (query) to IR (gallery) test mode (Red denotes the Top-1, while Green represents the Top-2). All the results are implemented based on our baseline. The training order is RegDB→SYSU-MM01→LLCM→HITSZ-VCM, thus $\bar{s}$ represents average adaptation performance on RegDB, SYSU-MM01, LLCM and HITSZ-VCM. Here, "Single-Learning" means training one dataset only using a single model, "Joint-Learning" represents training all datasets by a single model at one stage, while "Finetune" and other methods involve training all datasets one by one at each stage. Notably, in the first stage, in the interest of fairness, if a given method does not introduce additional novelty regarding the model's adaptability, we proceed by resuming directly from the model trained in the first stage using the Baseline, thereby rendering the majority of methods performance-equal in the first stage.

Experiment	Methods	RegDB (Stage 1)		SYSU-MM01 (Stage 2)		LLCM (Stage 3)		HITSZ-VCM (Stage 4)		$\bar{s}$ (Average)	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Comparison	Single-Learning	68.74	63.38	40.06	45.70	44.01	49.05	31.96	27.22	46.19	46.34
	Joint-Learning	72.77	71.87	49.82	52.73	56.33	60.57	40.97	34.40	54.97	54.89
	FineTune	68.74	63.38	42.15	46.79	47.59	52.69	28.21	25.09	46.67	46.99
	Lwf (Li and Hoiem, 2018)	68.74	63.38	38.72	43.57	44.23	50.20	14.31	13.95	41.50	42.78
	WA (Feng et al., 2023)	68.74	63.38	34.44	38.96	32.53	37.24	13.63	13.51	37.34	38.27
	BiC (Wu et al., 2019)	68.74	63.38	37.03	42.30	39.92	46.52	16.59	15.97	40.57	42.04
	DER (Buzzega et al., 2020)	68.74	63.38	33.51	37.69	34.07	39.85	26.84	23.57	40.79	41.12
	DER++ (Buzzega et al., 2020)	68.74	63.38	35.84	37.98	29.86	34.41	27.57	23.72	40.50	39.87
	AKA (Pu et al., 2021)	68.25	64.31	36.52	41.74	40.22	46.55	18.05	16.94	40.76	42.39
	PTKP (Ge et al., 2022)	63.69	63.43	37.82	41.92	41.00	46.46	30.02	25.96	43.13	44.44
	KRKC (Yu et al., 2023)	68.74	63.38	40.04	45.61	46.87	52.54	28.66	25.68	46.08	46.80
Ablation	Base.	68.74	63.38	36.94	41.33	45.93	50.90	30.26	26.14	45.47	45.44
	Base. + CDCN	83.74	78.62	43.06	46.67	50.03	54.47	34.15	29.16	52.75	52.23
	Base. + DMM	68.74	63.38	41.56	44.13	39.78	44.52	27.89	23.71	44.49	43.94
	Base. + CDCN + DMM	83.74	78.62	47.78	50.50	45.85	50.95	34.74	28.61	53.03	52.17

Table 2
Anti-forgetting performance under VIS (query) to IR (gallery) test mode at the last training stage (Red denotes the Top-1, while Green represents the Top-2). All the results are implemented based on our baseline. The training order is RegDB→SYSU-MM01→LLCM→HITSZ-VCM, hence $\bar{m}$ indicates weighted average performance on RegDB, SYSU-MM01, and LLCM after training HITSZ-VCM. Here, Q is the number of query samples per domain, with the cumulative query samples for computing $\bar{m}$ totaling 17515, consolidating the samples from the initial three datasets. The same as Table 1, "Single-Learning" means training one dataset only using a single model, "Joint-Learning" represents training all datasets by a single model at one stage, while "Finetune" and other methods involve training all datasets one by one at each stage.

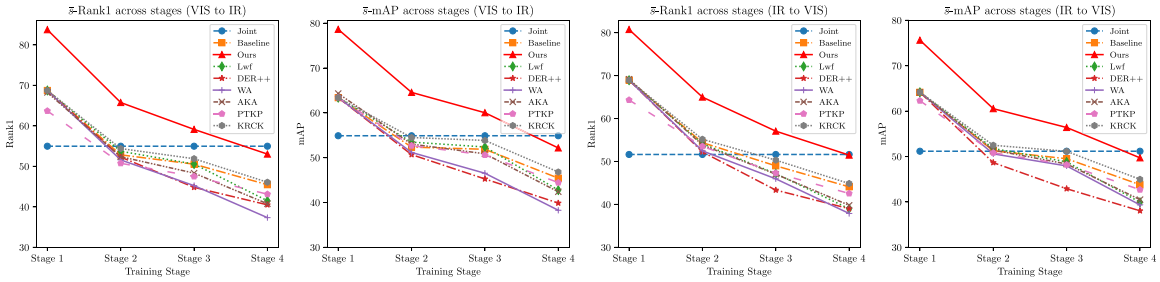
Experiment	Methods	RegDB (Q = 2060)		SYSU-MM01 (Q = 6775)		LLCM (Q = 8680)		$\bar{m}$ (Q = 17515)	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Comparison	Single-Learning	68.74	63.38	40.06	45.70	44.01	49.05	45.39	49.44
	Joint-Learning	72.77	71.87	49.82	52.73	56.33	60.57	55.75	58.87
	FineTune	7.09	8.26	11.35	15.93	11.23	14.97	10.79	14.55
	Lwf (Li and Hoiem, 2018)	49.13	45.72	20.37	24.48	21.67	26.42	24.40	27.94
	WA (Feng et al., 2023)	46.65	44.63	19.14	23.45	21.18	26.12	23.39	27.26
	BiC (Wu et al., 2019)	46.21	43.01	20.93	25.46	18.77	23.22	22.83	26.41
	DER (Buzzega et al., 2020)	28.54	28.00	28.25	32.61	36.39	41.59	32.32	36.52
	DER++ (Buzzega et al., 2020)	28.64	27.79	27.99	31.66	39.07	43.75	33.56	37.20
	AKA (Pu et al., 2021)	39.32	37.26	10.46	14.83	9.10	13.38	13.18	16.75
	PTKP (Ge et al., 2022)	49.08	47.45	31.20	35.49	32.17	37.66	33.78	37.97
	KRKC (Yu et al., 2023)	8.35	11.04	13.98	19.07	17.58	21.03	15.10	19.10
Ablation	Base.	43.40	42.01	29.33	32.38	36.75	41.99	34.66	38.28
	Base. + CDCN	56.84	54.19	28.92	33.66	39.76	44.23	37.58	41.31
	Base. + DMM	64.08	61.98	34.88	38.61	35.33	40.71	38.54	42.40
	Base. + CDCN + DMM	63.83	59.82	38.51	40.86	39.93	44.75	42.19	45.02

10078 infrared tracklets. In our experiments, we select 500 identities for training and 427 identities for the testing phase following (Lin et al., 2022) (the number of query samples is 121198).

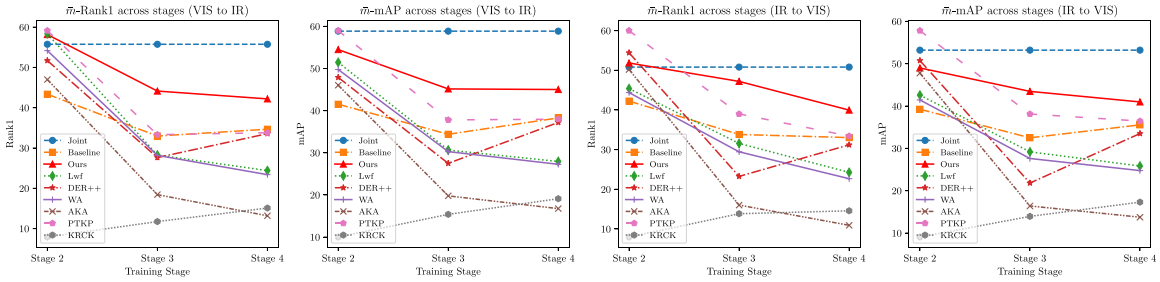
Given that the training set of the HITSZ-VCM dataset encompasses approximately 200000 images, a quantity significantly surpassing the combined total of the other three datasets, we adopt a tailored approach to mitigate potential performance degradation stemming from severe sample imbalance. Specifically, we randomly selected three frames from each tracklet within the HITSZ-VCM dataset to construct a new training set. The newly formed training set encompasses a total of 500 distinct identities and incorporates 16380 visible images alongside 12873 infrared images.

4.2. Implementation details

Following the previous VI-ReID works (Ye et al., 2021; Zhang et al., 2021), we adopt the two-stream ResNet-50 network as our backbone. Following the setting of AKA (Pu et al., 2021), all images are resized to 256 × 128, and we apply random horizontal flipping and random crop techniques to images during both the adaptation and rehearsal phases. For the experimental setup, we sequentially train on each data domain at each stage, dedicating 60 epochs for every stage. For the initial stage, the training process solely contains an adaptation phase. From the second stage onwards, each training epoch incorporates both an adaptation phase and a rehearsal phase. Specifically, in an adaptation mini-batch, we will sample 8 instances, each containing 4 visible images and 4 infrared images. In a rehearsal mini-batch, we will also randomly sample 8 instances from the memory bank, each containing 2 visible images and 2 infrared images. An SGD optimizer and a warm-up strategy (Luo et al., 2020) are used to optimize. The initial learning rate is set to 1×10^{-2}, then increases to 1×10^{-1} after 10 epochs. Thereafter, the learning rate is decayed by a factor of 0.1 at the 20th, 30th, and 40th epoch. At the first stage, the memory bank is empty. To update the memory bank, after the completion of training



(a) Adaptation performance $\bar{s}$ across stages, under both VIS to IR and IR to VIS test modes.



(b) Anti-forgetting performance across stages, under both VIS to IR and IR to VIS test modes. Notably, the calculation of $\bar{m}$ begins following the completion of the second training stage.

Fig. 4. Performances in adaptability 4(a) and anti-forgetting ability 4(b) under both VIS to IR and IR to VIS test modes. Our method achieves optimal stage-wise adaptation and enhanced long-term robustness in LVI-ReID.

at each stage, we randomly select 256 classes from current domain and employ the sample strategy in iCaRL (Rebuffi et al., 2017) to select 2 visible images and 2 infrared images of each class. Thus, we will store about 1024 samples after each stage, if the number of classes in a domain is less than 256, we select all classes from that domain. Notably, the memory bank will permanently retain stored samples. In our DMM strategy, the cross-domain mixing probability p_{cd} is set to 0.5, whereas the cross-modal mixing probability p_{cm} is set to 0.05.

4.3. Evaluation protocols

Following the evaluation protocol in LUDA-ReID (Huang et al., 2022), for the adaptation performance, we first evaluate models at the end of each stage, then calculate $\bar{s}$ (average adaptation performance on all seen domains) at the end of the last training stage to measure the comprehensive adaptability of models. For anti-forgetting performance, we employ $\bar{m}$ (weighted average performance on previously seen domains before the current training domain) to measure the anti-forgetting ability of models. Both $\bar{s}$ and $\bar{m}$ are measured with Rank-1 (R-1) accuracy and mean Average Precision (mAP).

4.4. Comparison to the state-of-the-art methods

We compared our proposed methods to several state-of-the-art methods, encompassing Lifelong Learning (LL) approaches: LwF (Li and Hoiem, 2018), DER/DER++ (Buzzega et al., 2020), WA (Feng et al., 2023), and BiC (Wu et al., 2019), as well as Lifelong Person Re-Identification (LReID) methods: AKA (Pu et al., 2021), PTKP (Ge et al., 2022), and KRKC (Yu et al., 2023). The adaptation and anti-forgetting performances of each method are presented in Table 1 and Table 2, respectively. The term "Single-Learning" denotes only training on a single dataset exclusively, whereas "Joint-Learning" refers to concurrent training across all datasets. "FineTune" represents the sequential training approach without any specialized retention strategies. For each of these compared methods, we meticulously implemented and fine-tuned them based on their published codes. To ensure a fair comparison, all methods were subjected to an identical backbone network architecture, preprocessing procedures, and training techniques consistent with those adopted in our method.

Adaptation Performance. Our method (Base. + CDCN + DMM) significantly outperforms other state-of-the-art approaches in the term of adaptability (in Table 1). Specifically, under the VIS to IR test mode, our method has achieved enhancements of 11.53 and 9.39 in Rank-1 and mAP, respectively, in $\bar{s}$ when compared to LL methods, while outperforming the Lifelong ReID methods by 6.95 Rank-1 and 5.37 mAP. Under the IR to VIS test mode (Table 3), our method achieves improvements of 6.62 Rank-1 and 4.73 mAP over the average adaptation performance $\bar{s}$ of LL and LReID approaches. This may be attributed to the fact that LL approaches are primarily designed for closed datasets, whereas LReID methods tend to lack considerations for cross-modal knowledge learning.

Anti-forgetting Performance. With regard to anti-forgetting ability, our method (Base. + CDCN + DMM) also demonstrates remarkable improvements (in Table 2). Under the VIS to IR test mode, our method demonstrates notable improvements compared to LL methods, achieving increases of 8.63 Rank-1 and 7.82 mAP, respectively, in $\bar{m}$. Furthermore, compared with LReID methods, our method surpasses them with enhancements of 8.41 Rank-1 and 7.05 mAP. While we achieve 6.54 Rank-1 and 4.52 mAP increases under the IR to VIS test mode (in Table 3). This observed superiority of our method likely arises from addressing the issue of catastrophic forgetting induced by the modal gap, and this deficiency is common to both LL and LReID approaches.

Stage-wise Performance. As depicted in Fig. 4, we evaluated the $\bar{s}$ and $\bar{m}$ metrics for each model at every stage. Regarding adaptability, our approach achieved the best performance across all stages. With respect to anti-forgetting capability, while our method did not yield the optimal results in the second phase, it went on to achieve the best performance in the long run as the training progressed through subsequent stages. This discrepancy in the second phase is likely attributed to the fact that, during this phase, the memory bank comprised data from a single domain only, precluding the mix of cross-domain images. Consequently, the model was unable to unleash its potential fully. Our model substantially boosts its anti-forgetting ability as training progresses, and the memory bank encompasses a more diverse dataset. This accentuates the efficacy of our approach in facilitating lifelong learning, underlining its prowess in sustaining the continuous acquisition of new knowledge without discarding previously learned information.

Table 3
Performances in adaptability $\bar{s}$ and anti-forgetting ability $\bar{m}$ under IR (query) to VIS (gallery) test mode. The training processes and evaluation procedures are the same as Table 1 and 2.

Methods	$\bar{s}$		$\bar{m}$	
	R-1	mAP	R-1	mAP
Joint-Learning	51.63	51.17	50.80	53.20
Lwf (Li and Hoiem, 2018)	39.00	39.94	24.26	25.83
WA (Feng et al., 2023)	37.89	39.35	22.67	24.75
DER++ (Buzzega et al., 2020)	39.01	38.05	31.20	33.50
AKA (Pu et al., 2021)	39.80	40.51	10.91	13.74
PTKP (Ge et al., 2022)	42.55	42.70	33.42	36.45
KRKC (Yu et al., 2023)	44.86	44.98	14.57	17.30
Base.	44.09	43.80	33.03	35.55
Base. + CDCN	51.00	49.43	36.25	38.09
Base. + DMM	44.29	42.76	37.59	39.52
Base. + CDCN + DMM	51.48	49.71	39.96	40.97

Table 4
Structure analysis for DMM strategy. The training order and evaluation procedures are the same as Table 1 and 2.

Adaptation	Rehearsal	$\bar{s}$		$\bar{m}$	
		R-1	mAP	R-1	mAP
✓	✓	49.71	49.19	34.54	37.22
✓		52.23	50.87	28.62	31.79
	✓	53.03	52.17	42.19	45.02

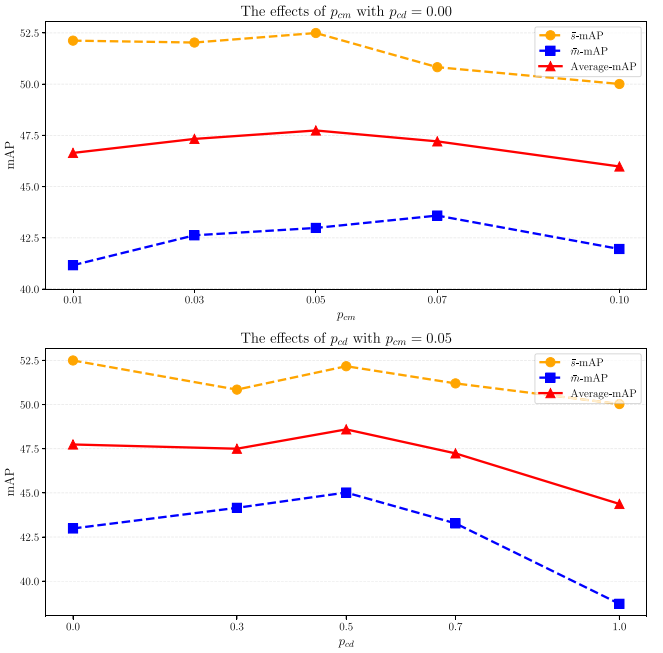


Fig. 5. Parameter Analysis of p_{cm} and p_{cd}. The Average-mAP indicates the mean performance of both $\bar{s}$ and $\bar{m}$.

4.5. Ablation studies

We performed ablation experiments to evaluate the individual impacts of distinct modules on the comprehensive enhancement of the model, as evidenced in Tables 1 and 2. Our research outcomes illustrate that integrating the replay sample Domain-Modal-Mixed (DMM) reconstruction strategy into the baseline model notably bolsters its resistance against catastrophic forgetting, although this comes with a trade-off of moderately compromising the model's adaptability. Simultaneously, incorporating the Cross-domain Cognitive Network (CDCN) results in a more conspicuous elevation of the model's adaptability throughout various stages, while concurrently bolstering its memory retention capabilities. The synergistic application of these two components yields a result where the whole exceeds the sum of its parts. The CDCN compensates for the adverse effect of DMM on adaptation, while DMM, in turn, reinforces the anti-forgetting memory efficacy of the CDCN.

4.5.1. The effectiveness of CDCN

As demonstrated in Table 1, following the integration of the CDCN module, the model (Base.+ CDCN) exhibits a notably enhanced adaptability across all stages compared to the baseline, with improvements including an increase of 15.00 Rank-1 and 15.24 mAP for RegDB, 6.12 Rank-1 and 5.34 mAP for SYSU-MM01, 4.1 Rank-1 and 3.57 mAP for LLCM, and 3.89 Rank-1 and 3.02 mAP for HITSZ-VCM. Meanwhile, compared to other methods, our CDCN achieves enhancements in the average adaptation performance $\bar{s}$, realizing a boost of 6.67 Rank-1 and 5.43 mAP under the VIS to IR test mode, and an increase of 6.14 Rank-1 and 4.45 mAP under the IR to VIS test mode (in Table 3), thereby demonstrating significant advancements in the cross-domain adaptability. Furthermore, our CDCN also facilitates an augmented resistance to forgetting in the model. With respect to the weighted average anti-forgetting performance $\bar{m}$, in comparison to other approaches, our method (Base.+CDCN) achieves an elevation of 3.8 Rank-1 and 3.34 mAP under the VIS to IR test mode, along with an enhancement of 2.83 Rank-1 and 1.64 mAP when tested under the inverse IR to VIS mode, thereby substantiating its efficacy in mitigating catastrophic forgetting across different modalities and domains learning tasks.

4.5.2. The effectiveness of DMM

After applying DMM to the baseline, regarding the weighted average anti-forgetting performance $\bar{m}$, in comparison to other approaches, our method yields enhancements of 4.76 Rank-1 and 4.43 mAP under the VIS to IR test mode (in Table 2), and further accomplishes increases of 4.17 Rank-1 and 3.07 mAP when tested under the IR to VIS mode (in Table 3), thereby validating its robust capacity in alleviating catastrophic forgetting in the LVI-ReID task.

Structure Analysis. To further evaluate the effectiveness of our DMM, we conducted comparative tests across three scenarios: adding mixed strategies solely during the adaptation phase, incorporating them only in the rehearsal phase, and applying mixed strategies in both the adaptation phase and the rehearsal phase, as detailed in Table 4. It has been observed that during the adaptation phase, the employment of a mixed strategy leads to a decline in the model's adaptability, potentially due to the CDCN's requirement to consolidate information from the same camera source when accruing knowledge. The implementation of the mixed strategy during the adaptation phase results in the accumulation of muddled knowledge, which not only diminishes the model's focus on learning the current domain but also disrupts previously learned knowledge, thereby severely undermining its anti-forgetting ability. Consequently, the exclusive application of the mixed strategy during the replay phase emerges as the optimal strategy.

Parameter Analysis. As depicted in Fig. 5, we employ hold-off validation data to ascertain two hyper-parameters p_{cm} and p_{cd} in the DMM. In order to strike an optimal balance between adaptability and anti-forgetting ability, we initially select p_{cm} to maximize the mean performance of both $\bar{s}$ and $\bar{m}$. Subsequently, based on the selected p_{cm}, we determine the optimal value for p_{cd}. Ultimately, our model achieves excellent performance in terms of both adaptability and anti-forgetting ability with $p_{cm} = 0.05$ and $p_{cd} = 0.5$.

4.5.3. The impacts of different training order

As shown Table 5, we test four different training order and found that when the fourth dataset is HITSZ-VCM, the model's performance fluctuates within a reasonable range (approximately 3 percentage points). At this point, despite different training orders, the experimental results do not show significant changes regardless of the order of the first three datasets. However, when the HITSZ-VCM dataset is

Table 5
Performance under different training orders. RE to represent the RegDB dataset, SY for SYSU-MM01, LL for LLCM, and HI for HITSZ-VCM datasets.

Training order	$\bar{s}$		$\bar{m}$	
	R-1	mAP	R-1	mAP
RE→SY→LL→HI	53.03	52.17	42.19	45.02
SY→RE→LL→HI	51.07	51.15	38.70	42.08
LL→SY→RE→HI	51.69	51.59	39.02	41.24
HI→LL→SY→RE	51.97	51.08	26.61	23.25

Table 6
Performance under different memory sizes. Notably, the memory size here represents the number of classes randomly selected from each domain. If the memory size is 256, it means that after each stage of training, we will select a total of 1024 samples (256 classes, each comprising 2 visible and 2 infrared samples). If the number of classes in a domain is less than 256, we select all classes from that domain.

Memory size	$\bar{s}$		$\bar{m}$	
	R-1	mAP	R-1	mAP
128	51.08	50.41	36.72	40.37
256	53.03	52.17	42.19	45.02
512	52.27	51.10	42.14	45.70
1024	49.78	49.21	40.42	43.33

trained first, there is a noticeable decline in the model's ability to resist forgetting. This might be caused by two factors: first, the HITSZ-VCM dataset is relatively complex, and even when trained last, as shown in Table 1, the model's best performance on the HITSZ-VCM dataset is relatively poor. Second, since the number of samples in the query set of HITSZ-VCM is much larger than in other datasets, the final $\bar{\boldsymbol{m}}$ are almost entirely determined by the HITSZ-VCM dataset, resulting in a large gap between the anti-forgetting performance and the basic level under this training order.

4.5.4. The impacts of different memory size

We test the impact of four levels of Memory Bank sizes, from small to large, on the final results, as shown in Table 6. When the memory size is too small (size = 128), the model's ability to resist forgetting shows a certain decline. However, when the memory is greater than or equal to 256, there is no significant change in the model's performance, and it remains within a reasonable fluctuation range.

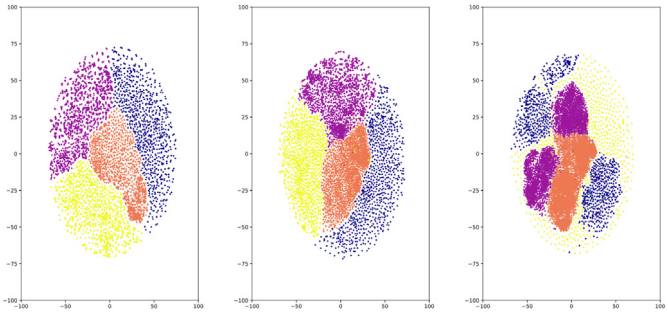
4.6. Visualization analysis

To provide a compelling visualization of the discriminative prowess of our methodology across multiple data domains, we randomly select 65536 samples (8192 visible samples and 8192 infrared samples from each domain). As depicted in Fig. 6, we conducted a comparative analysis involving our method, alongside Joint-Learning and FineTune strategies. It is revealed that the Finetune model nearly completely forfeits its discrimination when faced with data from different domains. In contrast, the discrimination of our method aligns closely with that of the Joint-Learning model, thus evidencing the strong potential of our proposed model in effectively distinguishing data from varied domains.

5. Discussion

5.1. Similarities and differences between CDCN and AKA

Our Cross-domain Cognitive Network (CDCN) and AKA (Pu et al., 2021) are similar in their LVI-ReID task application. In our batch-wise processing, akin to AKA, we adopt a fully connected framework, but we uniquely apply full connection operations separately to samples from the visible and infrared modalities. This tailored approach fosters learning of the specific consistencies within each modality, as opposed to improperly connecting all samples in a class batch without regard for their modal origins. A pivotal distinction sets our work apart from AKA.



(a) Joint (b) Ours (c) Finetune

Fig. 6. Comparative visualization of extracted features. From each dataset, we randomly selected 8192 infrared and 8192 visible images, with points of the same color representing the same dataset. Specifically, (a) displays features decompressed by the Joint-Learning model, (b) showcases those from our proposed model, and (c) presents features resulting from the Finetune model.

AKA employs a predefined, fixed-size memory network and connects it in a fully connected fashion to every sample within a batch, which may perform poorly in the face of profound inter-modal discrepancies. In contrast, our methodology integrates an adaptable, camera-aware network, which innovatively tackles modal disparity at the granularity of distinct cameras. By integrating cross-modal knowledge across multiple cameras within the domain, our CDCN mitigates the effects of modal differences more efficaciously. Furthermore, incorporating a shared GCN layer bolsters the model's capability to recall and solidify previously learned knowledge. In summary, while both methods are designed for the lifelong learning task, our CDCN tailors its fully connected processing for individual modalities and introduces a dynamic, adaptive network architecture specifically designed to handle cross-modal data originating from different domains.

5.2. The respective pros and cons of DMM and CDCN

How to balance adaptability and anti-forgetting ability in lifelong learning is a concern that has been of ongoing interest in academia. Our DMM is designed to tackle the reduction in anti-forgetting performance caused by limited replay samples which is equivalent to improving its anti-forgetting ability. Just like a student, If you help him review more varied prior knowledge, he may he may get distracted from learning new knowledge to some extent, so DMM alone can improve the anti-forgetting ability while cause a negative impact on adaptability. As for the CDCN, it is designed to help model adapt new knowledge mainly, we draw inspirations from chunking theory which divide knowledge into several aspects for better learning and design the cognitive nodes to accumulate chunking knowledge in the camera level, finally we use the cognitive nodes as anchors and the GCN layer as the lever to boost the adaptation of our model. However, due to the knowledge accumulated across domains by cognitive nodes, CDCN also becomes a extend memory module to improve the anti-forgetting ability of the model.

5.3. The limitations of our work

Despite the remarkable progress achieved by our proposed DMM in combating catastrophic forgetting, further exploration into the potential of mixing samples across domains and modalities remains a valuable issue. In stage-wise experiments (in Fig. 4), our model did not yield the best performance in short-term learning scenarios, which can potentially be attributed to an inability to effectively integrate cross-domain images. Additionally, applying too strong a DMM (by increasing the values of the mixing strategies p_{cm} and p_{cd}) may cause the model to focus most of its attention on revising old knowledge,

leading to difficulties in learning new knowledge. If the model cannot learn the current new knowledge well, it will still not perform well on this part of the knowledge once it becomes old, which is similar to human learning patterns. These aspects underscore the necessity for further investigation and refinement.

Meanwhile, Due to the scarcity of existing VI-ReID datasets, there is a limitation in thoroughly evaluating the extent to which our model can generalize to unseen datasets. Consequently, the comprehensive assessment of its generalization ability remains an open issue that requires further investigation with additional data resources.

To further improve LVI-ReID models, we propose three key directions: first, enhancing the model's ability to learn cross-modal images, as only by learning new knowledge well can the model more likely retain it when it becomes old; second, developing stronger replay sample cross-modal and cross-domain augmentation strategies to help the model achieve broader application when revisiting old knowledge; third, extracting more refined knowledge points for the model, akin to Prompts in Transformers, enabling it to pinpoint learned knowledge more accurately. In summary, some of the learning patterns observed in human lifelong learning can be applied to LVI-ReID to help the model perform better.

6. Conclusion

In this paper, we address the Lifelong Visible–Infrared Person ReIdentification (LVI-ReID) problem from a more fine-grained perspective, aiming to solve the critical issues of catastrophic forgetting and poor adaptability to new tasks encountered by models when learning cross-modal and cross-domain data. To effectively mitigate the detrimental effects of catastrophic forgetting, we introduce the Replay Sample Domain-Modal-Mix Reconstruction strategy to diversify the replayed data, thereby enhancing the model's resilience against memory decay caused by cross-modal and cross-domain differences. Concurrently, we propose a Camera-aware Cross-domain Cognitive Network that encourages the model to learn both intra-modal consistencies and cross-modal similarities. The proposed methods not only enhance the model's adaptability but also significantly reinforce its retention of previously acquired knowledge, thus endowing it with improved long-term learning performance in the context of visible–infrared person re-identification tasks.

CRediT authorship contribution statement

Xianyu Zhu: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Guoqiang Xiao:** Writing – review & editing. **Michael S. Lew:** Writing – review & editing. **Song Wu:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Sichuan Science and Technology Program, China (2025ZNSFSC0482).

Data availability

Data will be made available on request.

References

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T., 2018. Memory aware synapses: Learning what (not) to forget. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Cham, pp. 144–161.

Bai, Z., Wang, Z., Wang, J., Hu, D., Ding, E., 2021. Unsupervised multi-source domain adaptation for person re-identification. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 12909–12918.

Bang, J., Kim, H., Yoo, Y., Ha, J.-W., Choi, J., 2021. Rainbow memory: Continual learning with a memory of diverse samples. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 8214–8223.

Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S., 2020. Dark experience for general continual learning: a strong, simple baseline. arXiv abs/2004.07211.

Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C., 2020. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 10254–10263.

Feng, Y., Zhao, P., Guo, Y., Zhao, X., Hao, W., Zhao, R., Li, F., 2023. Maintaining distinction and fairness in data-free class incremental learning. In: 2023 International Joint Conference on Neural Networks. IJCNN, pp. 1–7.

Ge, W., Du, J., Wu, A., Xian, Y., Yan, K., Huang, F., Zheng, W.-S., 2022. Lifelong person re-identification by pseudo task knowledge preservation. Proc. the AAAI Conf. Artif. Intell. 36 (1), 688–696.

Gobet, F., Lane, P.C., Croker, S., Cheng, P.C.-H., Jones, G., Oliver, I., Pine, J.M., 2001. Chunking mechanisms in human learning. Trends Cogn. Sci. 5 (6), 236–243.

Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D., 2019. Learning a unified classifier incrementally via rebalancing. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 831–839.

Hua, B., Zhang, J., Li, Z., Ge, Y., 2023. Cross-modality channel mixup and modality decorrelation for RGB-infrared person re-identification. IEEE Trans. Biom. Behav. Identity Sci. 5 (4), 512–523.

Huang, Z., Zhang, Z., Lan, C., Zeng, W., Chu, P., You, Q., Wang, J., Liu, Z., Zha, Z.-J., 2022. Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 14268–14277.

Jiang, B., Zhang, Z., Lin, D., Tang, J., Luo, B., 2019. Semi-supervised learning with graph learning-convolutional networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 11305–11312.

Kansal, K., Subramanyam, A.V., Wang, Z., Satoh, S., 2020. SDL: Spectrum-disentangled representation learning for visible-infrared person re-identification. IEEE Trans. Circuits Syst. Video Technol. 30 (10), 3422–3432. http://dx.doi.org/10.1109/TCSVT.2019.2963721.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N.C., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R., 2016. Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. 114, 3521–3526.

Kong, Y., Liu, L., Qiao, M., Wang, Z., Tao, D., 2023. Trust-region adaptive frequency for online continual learning. Int. J. Comput. Vis. 131 (7), 1825–1839.

Li, Z., Hoiem, D., 2018. Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. 40 (12), 2935–2947.

Li, D., Wei, X., Hong, X., Gong, Y., 2020. Infrared-visible cross-modal person re-identification with an X modality. Proc. the AAAI Conf. Artif. Intell. 34 (04), 4610–4617.

Lin, X., Li, J., Ma, Z., Li, H., Li, S., Xu, K., Lu, G., Zhang, D., 2022. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 20941–20950.

Liu, J., Sun, Y., Zhu, F., Pei, H., Yang, Y., Li, W., 2022. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 19344–19353.

Liu, S., Xiao, G., Lew, M.S., Gao, X., Wu, S., 2024. Core-attributes enhanced generative adversarial networks for robust image enhancement. Eng. Appl. Artif. Intell. 131, 107799. http://dx.doi.org/10.1016/j.engappai.2023.107799.

Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N., 2020. Cross-modality person re-identification with shared-specific feature transfer. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 13376–13386.

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J., 2020. A strong baseline and batch normalization neck for deep person re-identification. IEEE Trans. Multimed. 22 (10), 2597–2609.

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

Pu, N., Chen, W., Liu, Y., Bakker, E.M., Lew, M.S., 2021. Lifelong person re-identification via adaptive knowledge accumulation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 7897–7906.

Rannen, A., Aljundi, R., Blaschko, M.B., Tuytelaars, T., 2017. Encoder based lifelong learning. In: 2017 IEEE International Conference on Computer Vision. ICCV, pp. 1329–1337.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., Lampert, C.H., 2017. iCaRL: Incremental classifier and representation learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5533–5542.

Serra, J., Suris, D., Miron, M., Karatzoglou, A., 2018. Overcoming catastrophic forgetting with hard attention to the task. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 80, pp. 4548–4557.

Shen, Y., Li, H., Yi, S., Chen, D., Wang, X., 2018. Person re-identification with deep similarity-guided graph neural network. In: Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV. Berlin, Heidelberg, pp. 508–526.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S., 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Cham, pp. 501–518.

Vitter, J.S., 1985. Random sampling with a reservoir. ACM Trans. Math. Software 11 (1), 37–57.

Wang, Z., Liu, L., Duan, Y., Kong, Y., Tao, D., 2022. Continual learning with lifelong vision transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 171–181.

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

Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z., 2019b. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In: 2019 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 3622–3631.

Wei, Z., Yang, X., Wang, N., Gao, X., 2021. Syncretic modality collaborative learning for visible infrared person re-identification. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 225–234.

Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y., 2019. Large scale incremental learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 374–382.

Wu, G., Gong, S., 2021. Generalising without forgetting for lifelong person re-identification. Proc. the AAAI Conf. Artif. Intell. 35 (4), 2889–2897.

Wu, S., Shan, S., Xiao, G., Lew, M.S., Gao, X., 2024a. Modality blur and batch alignment learning for twin noisy labels-based visible–infrared person re-identification. Eng. Appl. Artif. Intell. 133, 107990. http://dx.doi.org/10.1016/j.engappai.2024.107990.

Wu, S., Yuan, X., Xiao, G., Lew, M.S., Gao, X., 2024b. Deep cross-modal hashing with multi-task latent space learning. Eng. Appl. Artif. Intell. 136, 108944. http://dx.doi.org/10.1016/j.engappai.2024.108944.

Wu, A., Zheng, W.-S., Gong, S., Lai, J., 2020. RGB-IR person re-identification by cross-modality similarity preservation. Int. J. Comput. Vis. 128 (6), 1765–1785.

Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S., Lai, J., 2017a. RGB-infrared cross-modality person re-identification. In: 2017 IEEE International Conference on Computer Vision. ICCV, pp. 5390–5399.

Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S., Lai, J., 2017b. RGB-infrared cross-modality person re-identification. In: 2017 IEEE International Conference on Computer Vision. ICCV, pp. 5390–5399.

Xing, Y., Xiao, G., Lew, M.S., 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. ICMR '24, Association for Computing Machinery, New York, NY, USA, pp. 988–997. http://dx.doi.org/10.1145/3652583.3658033.

Xu, B., Wang, N., Chen, T., Li, M., 2015. Empirical evaluation of rectified activations in convolutional network. CoRR abs/1505.00853.

Ye, M., Ruan, W., Du, B., Shou, M.Z., 2021. Channel augmented joint learning for visible-infrared recognition. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 13547–13556.

Ye, M., Shen, J., J. Crandall, D., Shao, L., 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. Berlin, Heidelberg, pp. 229–247. http://dx.doi.org/10.1007/978-3-030-58520-4_14.

Yoon, J., Yang, E., Lee, J., Hwang, S.J., 2017. Lifelong learning with dynamically expandable networks. CoRR abs/1708.01547.

Yu, C., Shi, Y., Liu, Z., Gao, S., Wang, J., 2023. Lifelong person re-identification via knowledge refreshing and consolidation. In: AAAI Conference on Artificial Intelligence. In: AAAI'23/IAAI'23/EAAI'23, vol. 37, (no. 3), pp. 3295–3303.

Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J., 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In: 2019 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 6022–6031.

Zhang, Y., 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. CVPR, pp. 2153–2162.

Zhang, Y., Yan, Y., Lu, Y., Wang, H., 2021. Towards a unified middle modality learning for visible-infrared person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia. MM '21, New York, NY, USA, pp. 788–796.

Zhong, X., Lu, T., Huang, W., Ye, M., Jia, X., Lin, C.-W., 2022. Grayscale enhancement colorization network for visible-infrared person re-identification. IEEE Trans. Circuits Syst. Video Technol. 32 (3), 1418–1430. http://dx.doi.org/10.1109/TCSVT.2021.3072171.

Zhu, J., Luo, B., Zhao, S., Ying, S., Zhao, X., Gao, Y., 2020. IExpressNet: Facial expression recognition with incremental classes. In: Proceedings of the 28th ACM International Conference on Multimedia. MM '20, New York, NY, USA, pp. 2899–2908.