Domain Consistency Representation Learning for Lifelong Person Re-Identification

Shiben Liu, Huijie Fan, Qiang Wang, Weihong Ren, Yandong Tang, Yang Cong, *Senior Member, IEEE*

arXiv:2409.19954v4 [cs.CV] 3 Sep 2025

***Abstract*—Lifelong person re-identification (LReID) exhibits a contradictory relationship between intra-domain discrimination and inter-domain gaps when learning from continuous data. Intra-domain discrimination focuses on individual nuances (*i.e.*, clothing type, accessories, *etc.*), while inter-domain gaps emphasize domain consistency. Achieving a trade-off between maximizing intra-domain discrimination and minimizing inter-domain gaps is a crucial challenge for improving LReID performance. Most existing methods strive to reduce inter-domain gaps through knowledge distillation to maintain domain consistency. However, they often ignore intra-domain discrimination. To address this challenge, we propose a novel domain consistency representation learning (DCR) model that explores global and attribute-wise representations as a bridge to balance intra-domain discrimination and inter-domain gaps. At the intra-domain level, we explore the complementary relationship between global and attribute-wise representations to improve discrimination among similar identities. Excessive learning intra-domain discrimination can lead to catastrophic forgetting. We further develop an attribute-oriented anti-forgetting (AF) strategy that explores attribute-wise representations to enhance inter-domain consistency, and propose a knowledge consolidation (KC) strategy to facilitate knowledge transfer. Extensive experiments show that our DCR achieves superior performance compared to state-of-the-art LReID methods. Our code is available at https://github.com/LiuShiBen/DCR.**

***Index Terms*—Lifelong person re-identification, attribute-text generator, text-image aggregation, domain consistency representation.**

I. Introduction

PERSON re-identification (ReID) aims to retrieve the same individual across multiple cameras in a large-scale database by using uni-modal architectures such as convolutional neural networks (CNN) [1]–[3] or vision transformers (ViT) [4]–[6]. However, when ReID models are applied to continuous datasets collected by video-based monitoring systems [7], [8], they exhibit notable performance limitations. As a result, recent works have focused on the practical problem of lifelong person identification (LReID), which maintains strong performance with continuously updated data streams.

At present, lifelong person re-identification (LReID) suffers from the challenge of balancing the anti-forgetting of old knowledge and learning new knowledge. Specifically, there are two main issues to solve this challenge. 1) **Intra-domain discrimination**. Each identity may exhibit subtle nuances of individual information (*i.e.*, clothing type, accessories, haircut, *etc.*) and lead to severe distribution overlapping. Learning discriminative representations of individuals are effective for distinguish identity information. 2) **Inter-domain gaps**. Each

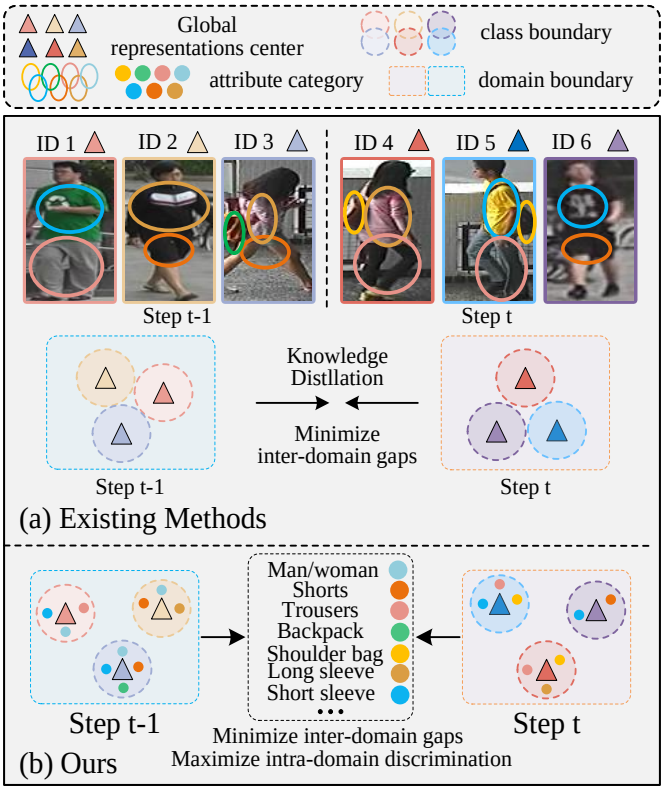


Fig. 1. Comparison between our method and existing methods. (a) Existing methods [9], [10] leverage knowledge distillation to minimize inter-domain gaps but ignore intra-domain discrimination, which limits the LReID model's ability to learn new knowledge. (b) Our method explores domain consistency representations as a bridge to achieve a trade-off between maximizing intra-domain discrimination and minimizing inter-domain gaps, enhancing the LReID model's anti-forgetting and generalization capabilities.

This work is supported by the National Natural Science Foundation of China (62273339, U24A201397), the Key Research and Development Program of Liaoning (2024JH2/102400022) and the LiaoNing Revitalization Talents Program (XLYC2403128). (*Corresponding author: Huijie Fan*)

Shiben Liu is with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: liushiben@sia.cn).

Huijie Fan, and Yandong Tang are with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China (e-mail: fanhuijie@sia.cn; ytang@sia.cn).

Qiang Wang is with the Key Laboratory of Manufacturing Industrial Integrated Automation, Shenyang University, and with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China (e-mail: wangqiang@sia.cn).

Weihong Ren is with the Harbin Institute of Technology, Shenzhen 518055, China (e-mail: renweihong@hit.edu.cn).

Yang Cong is with the College of Automation Science and Engineering, South China University of Technology, Guangzhou, 510640, China (e-mail: congyang81@gmail.com).

Domain is collected in different illumination and background, leading to inter-domain gaps. Bridging intra-domain gaps are significant for mitigating catastrophic forgetting in LReID.

To address these issues, we aim to learn domain consistency representations that capture individual nuances in intra-domain and inter-domain consistency in LReID. Knowledge distillation-based approaches [10]–[12] ensure distribution consistency between the previous and current domain to alleviate catastrophic forgetting. However, these approaches impose strict constraints and ignore intra-domain discrimination, [13]–[15], as outlined in Fig. 1(a). While LReID models significantly improve intra-domain discrimination for the current step, they inevitably damage inter-domain consistency, leading to catastrophic forgetting. Thus, we explore global and attribute-wise representations to strike a trade-off between maximizing intra-domain discrimination and minimizing inter-domain gaps, improving the anti-forgetting and generalization capabilities of the LReID model, as illustrated in Fig. 1(b).

Specifically, we propose a novel domain consistency representation learning (DCR) model that first explores attribute and text information to enhance LReID performance. Unlike methods [16]–[18], we develop domain consistency representations including global and attribute-wise representations to capture individual nuances in intra-domain and inter-domain consistency in LReID. We design an attribute-text generator (ATG) to dynamically generate text-image pairs for each instance, which are then fed into a text-guided aggregation (TGA) network to enhance the global representation capability, effectively distinguishing identities in LReID. Furthermore, the attributes of each instance guide an attribute compensation (ACN) network to generate attribute-wise representations that focus on specific regional information about identities. We consider that attributes can enhance reliability by setting higher thresholds across domains. Therefore, the generated attribute-wise representations and text for each instance are considered reliable in our model.

In summary, we aim to strike a balance between maximizing intra-domain identity-discriminative information and minimizing inter-domain gaps by exploring global and attribute-wise representations. At the intra-domain level, global representations capture whole-body information, while attribute-wise representations focus on specific regional information. When whole-body appearances or attribute-related information are similar across identities, we combine global and attribute-wise representations to distinguish among similar identities, maximizing intra-domain discrimination. While the model adapts well to new information, it often struggles with catastrophic forgetting. To preserve old knowledge, we develop an attribute-oriented anti-forgetting (AF) strategy that explores attribute-wise representations to bridge inter-domain gaps across continuous datasets. Meanwhile, striking a trade-off between maximizing intra-domain discrimination and minimizing inter-domain gaps is crucial for improving the performance of LReID models. Knowledge consolidation (KC) is proposed to facilitate knowledge transfer and enhance generalization capabilities, which consists of alignment and logit-level distillation mechanisms. The alignment mechanism explores global representations of knowledge transfer from the current dataset. The logit-level distillation mechanism enhances the extraction of identity information, thereby improving the model's ability to consolidate knowledge. Our contributions are as follows:

- We propose a novel domain consistency representation learning (DCR) model that explores global and attribute-wise representations to capture individual nuances in intra-domain and inter-domain consistency, achieving a trade-off between maximizing intra-domain discrimination and minimizing inter-domain gaps.
- In the intra-domain context, we explore the complementary relationship between global and attribute-wise representations to enhance the discrimination of each identity and adapt to new knowledge.
- In the inter-domain context, we design an attribute-oriented anti-forgetting (AF) and a knowledge consolidation (KC) strategy to minimize inter-domain gaps and facilitate knowledge transfer, improving the LReID model's generalization and anti-forgetting capabilities.

II. Related work

A. Lifelong Person Re-Identification

Lifelong Person Re-Identification (LReID) aims to balance intra-domain discrimination with minimizing inter-domain gaps in continuously updated datasets across scenarios, improving the model's anti-forgetting and generalization capabilities. LReID methods can be divided into two categories. 1) *Knowledge distillation-based methods* [12], [19]–[21] utilize metric strategies to achieve domain-consistent alignment between the old model with learned knowledge distribution and the new model that adaptively learns new knowledge. 2) *Exemplar-based methods* [9], [10], [22] achieve a distribution balance between old and new samples to prevent catastrophic forgetting by forming a memory buffer to select the limited samples from some identities. These methods strive to reduce inter-domain gaps and ensure consistency across domains to prevent catastrophic forgetting. However, they ignore intra-domain identity discrimination and lack consistency optimization within the inter-domain, limiting the LReID model's performance in learning new knowledge. In this paper, we explore domain consistency representations as a bridge to achieve a trade-off between maximizing intra-domain discrimination and minimizing inter-domain gaps for enhancing the anti-forgetting and generalization capabilities of the LReID model.

B. Vision-Language for Person Re-Identification

Vision-language learning paradigms [23], [24] have gained widespread popularity in recent years. Contrastive Language-Image Pre-training (CLIP) [25], establishes a connection between natural language and visual content through the similarity constraint of image-text pairs. CLIP has been applied to multiple person re-identification tasks [26]–[28], including text-to-image, text-based single-modality, and text-based

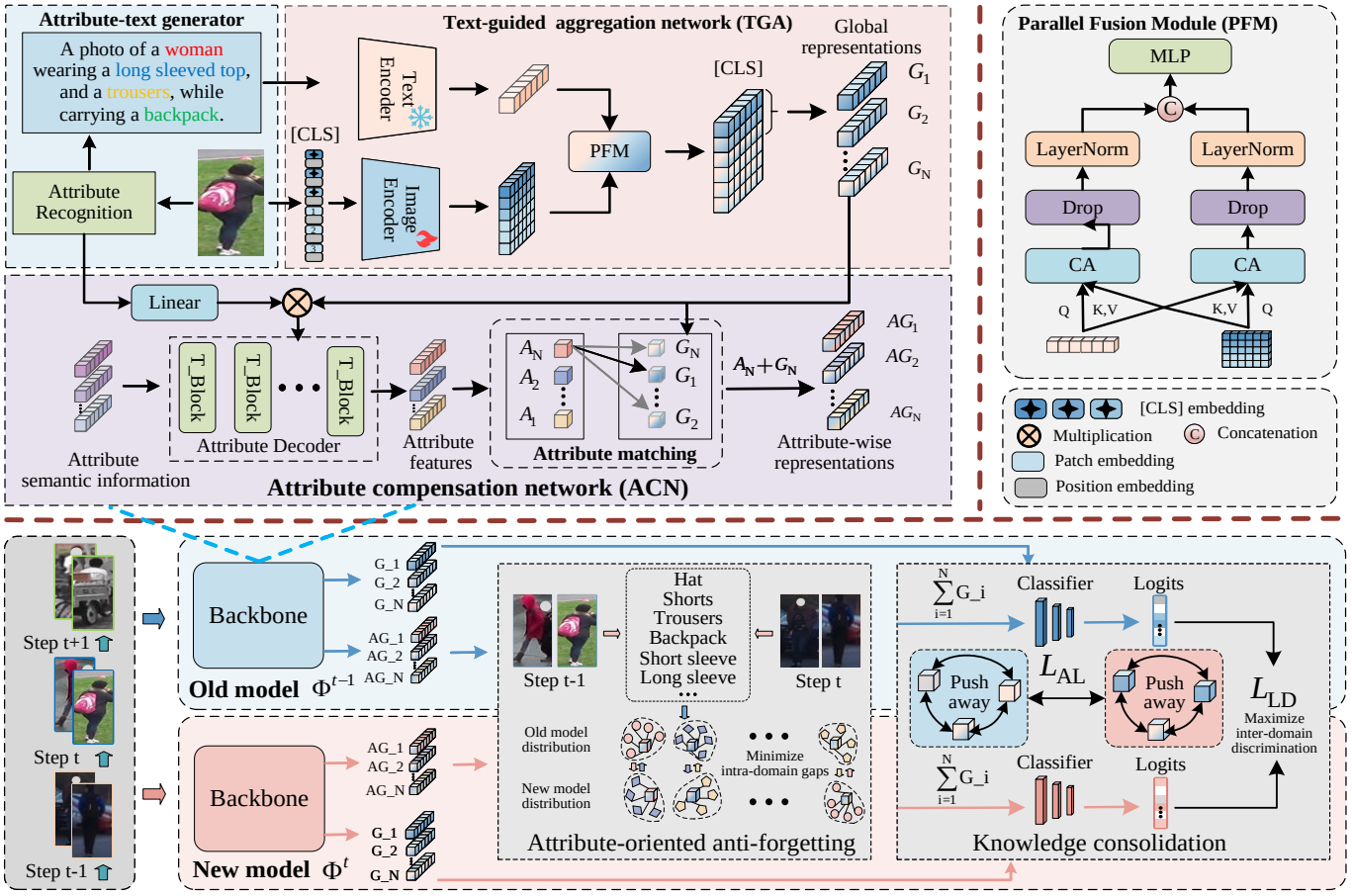


Fig. 2. Overview of the proposed DCR for LReID. First, the attribute-text generator (ATG) dynamically generates text-image pairs for each instance. Then, the text-guided aggregation network (TGA) captures global representations for each identity, while the attribute compensation network (ACN) generates attribute-wise representations. We explore the complementary relationship between global and attribute-wise representations to maximize intra-domain discrimination. Meanwhile, we design attribute-oriented anti-forgetting (AF) and knowledge consolidation (KC) strategies to minimize inter-domain gaps and facilitate knowledge transfer.

cross-modality. Text-to-image methods [28]–[30] aim to retrieve the target person based on a textual query. Text-based single-modality works [5], [27], [31] leverage text descriptions to generate robust visual features or integrate the beneficial features of text and images for the person category. Text-based cross-modality methods [32] utilize text descriptions to reduce visible-infrared modality gaps. Providing insufficient text descriptions of each identity, due to prompt learning and text inversion. In this paper, we dynamically generate text-image pairs from single images to capture fine-grained global representations based on the CLIP model for improving model performance capability in terms of inter-domain and intra-domain.

C. Pedestrian Attribute Recognition

Pedestrian attribute recognition aims to assign a set of attributes (Gender, Bag, Short/Long sleeve, and *etc.*) to a visual representation of a pedestrian based on their attributes. Deep learning-based research [33] automatically learns hierarchical features from raw images, improving recognition accuracy. Multi-task learning methods [34]–[36] leverage additional contextual information from multiple domains, such as pedestrian detection or pose estimation, to significantly improve attribute recognition. Part-based methods [37], [38] divide the pedestrian image into several parts or regions, providing more accurate localization. Currently, these methods have achieved significant success in improving the accuracy of attribute recognition. We are the first to explore the application of attributes to LReID from two perspectives. 1) Attributes are converted into text descriptions for each image to enhance global representation capabilities. 2) The attributes are transformed into attribute-wise representations by specific networks to maximize intra-domain discrimination and minimize intra-domain gaps.

III. Proposed Method

A. Preliminary: Overview of Method

The overview of our DCR to achieve a trade-off between maximizing intra-domain discrimination and inter-domain gaps is shown in Fig. 2. The DCR model learns the old model Φ^{t-1} and new model Φ^{t} from (t-1)-th and t-th steps, where Φ^{t} is inherited from Φ^{t-1}. Φ^{t-1} and Φ^{t} with three parts of attribute-text generator (ATG), text-guided aggregation

TABLE I
THE ATTRIBUTE CATEGORIES ARE DIVIDED INTO FOUR MAJOR CATEGORIES, INCLUDING 12 SUBCATEGORIES. "NO" AND "YES" INDICATE THE ABSENCE AND PRESENCE OF THIS CATEGORY, RESPECTIVELY.

Type	Category	0	1
Overall	Sex	man	woman
Upper body	Short sleeved top	No	Yes
	Long sleeved top	No	Yes
	Long coat	No	Yes
Lower body	Trousers	No	Yes
	Shorts	No	Yes
	Skirt	No	Yes
Decoration	Hat	No	Yes
	Glasses	No	Yes
	Handbag	No	Yes
	Sholder bag	No	Yes
	Backpack	No	Yes

network (TGA), and attribute compensation network (ACN). ϕ^{t-1} and ϕ^t serve as classifier heads for the old and new models, providing logits of each instance for recognition. Additionally, we define that consecutive T person datasets $D = \{D^t\}_{t=1}^{T}$ are collected from different environments, and establish a memory buffer M to store a limited number of samples from each previous ReID dataset. Given an image $x_i^t \in \mathcal{D}^t \cup \mathcal{M}$, we forward it to Φ^{t-1} and Φ^t is as follows:

$$G^{t-1}, AG^{t-1} = \Phi^{t-1}(x_i^t); \quad G^t, AG^t = \Phi^t(x_i^t). \tag{1}$$

Where G and AG are global and attribute-wise representations, respectively.

B. Attribute-Text Generator

Due to the lack of text-image pairs in ReID datasets, we propose an attribute-text generator (ATG) to dynamically generate text-image pairs for each instance, as illustrated in Fig. 2. Firstly, we employ a pre-trained attribute recognition model [33] (trained on the PA100K dataset [39]) to generate attribute categories. These categories are organized into four major types (e.g., overall, upper body, lower body, and decoration) with 12 subcategories, as shown in TABLE I. Then, attribute categories are converted into text descriptions for each instance using a predefined template. Finally, the generated text-image pairs are fed into the text-guided aggregation network (TGA) for further processing. In this predefined template (Fig. 2), black font denotes fixed modifiers, while red, blue, orange, and green fonts represent the four major types, respectively. The template dynamically replaces the colored placeholders with the corresponding predicted attributes. To ensure robustness across domains, we set a high confidence threshold (0.80) for attribute classification, thereby maintaining the reliability of text descriptions despite potential domain variations.

C. Text-Guided Aggregation Network

To enhance robust representation capabilities for each instance, we propose a text-guided aggregation network (TGA) to explore the global representation for distinguishing identity information in LReID, as shown in Fig. 2 (TGA). The TGA consists of a Contrastive Language–Image Pretraining(CLIP) model and a parallel fusion module (PFM). Note that the text encoder is frozen in our DCR model.

1) CLIP Model: The CLIP model is a vision-language model that learns multi-modal representations by aligning images and text in a shared embedding space. The model employs a contrastive learning objective to maximize the similarity between an image and its corresponding text description. In our work, we adapt the CLIP model, which includes an image encoder and a text encoder, to generate both image and text embeddings. However, unlike the original CLIP model [25], our model introduces multiple [CLS] embeddings into the image encoder's input sequence. This modification enables the model to capture diverse representations from different views for each instance, thereby enhancing its ability to distinguish fine-grained identity information. Specifically, the clip model generate text embedding d^* and image embedding $[v_1^*, \cdots, v_N^*, v_1, \cdots, v_P]$ for extracting rich text and vision information. image embedding consists of multiple [CLS] embeddings $[v_1^*, \cdots, v_N^*]$ for multi-view representations and patch embedding $[v_1, \cdots, v_P]$ for local feature extraction.

2) Parallel Fusion Module: To improve the performance of the LReID model, we propose a parallel fusion module (PFM) to explicitly explore the interactions between image and text embeddings, as shown in Fig. 2 (PFM). Firstly, we leverage text embedding d^* as query and image embedding $[v_1^*, \cdots, v_N^*, v_1, \cdots, v_P]$ as key and value to implement operation with cross-attention, drop, and layer normalization, getting text-wise representations. Similarly, in another fusion branch, image-wise representations are obtained. Finally, image-wise and text-wise representations perform concatenation and MLP operations to obtain global representations $G^t = \{G_i^t | i = 1, 2, \cdots, N\}$, focusing on whole body information. We force multiple global representations G^t at the current step to learn more discriminative information by orthogonal loss to minimize the overlapping elements. The orthogonal loss can be expressed as:

$$L_{Ort} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (G_i^t, G_j^t). \tag{2}$$

Then, we utilize the cross-entropy loss L_{CE} and triplet loss L_{Tri}^g [5] to optimize our DCR at the current task.

$$L_{\mathrm{CE}} = \frac{1}{K} \sum_{i=1}^{K} y_i \log((\phi^t(G^t))_i). \tag{3}$$

$$L_{\mathrm{Tri}}^g = max(d_p^g - d_n^g + m, 0). \tag{4}$$

Where K is the number of classes, and m is the margin, d_p^g and d_n^g are the distances from positive samples and negative samples to anchor samples in global representations, respectively. Unlike some methods [10], [13], global representations generated by the text-guided aggregation (TGA) network present two advantages. First, we leverage text descriptions based on the CLIP model to enhance the discrimination capability of global representations, allowing them to better

distinguish identities and adapt to new knowledge. Second, global representations facilitate knowledge transfer, improving the model's generalization ability.

D. Attribute Compensation Network

We force attributes to guide the attribute compensation network (ACN) for learning attribute-wise representations. The ACN consists of an attribute decoder and an attribute matching component, as illustrated in Fig. 2 (ACN).

1) Attribute Decoder: Enabling attributes to better adapt across domains, we define multiple learnable attributes semantic information $S^t = \{S_i^t | i = 1, 2, \cdots, N\}$ to learn discriminative information. The attributes undergo a linear layer to increase their dimensions and are then multiplied with the text-image global representation to output f_{AT}. Attribute semantic information S^t as queries Q, f_{AT} as keys and values are input into the attribute decoder, which outputs the attribute features $A^t = \{A_i^t | i = 1, 2, \cdots, N\}$. The attribute decoder utilizes six transformer blocks (T_Block) referenced from [40].

2) Attribute Matching: The attribute features $A^t = \{A_i^t | i = 1, 2, \cdots, N\}$ focus on multiple discriminative local details of individuals. However, it remains unclear which views make these attribute features more prominent in the global representations. We propose an Attribute Matching (AM) mechanism to select the optimal alignment between attribute features and global representations, thereby identifying the best view for the global representation and improving the representational ability of the attribute features. Specifically, we first calculate the similarity between the attribute features $A^t = \{A_i^t | i = 1, 2, \cdots, N\}$ and global representations $G^t = \{G_i^t | i = 1, 2, \cdots, N\}$ from different views. Based on this matching process, we determine the index of the most similar global representation as follows.

$$k = argmax(< A_i^t, G^t >). \tag{5}$$

where $<, >$ represent cosine similarity. The $argmax$ operation selects both the maximum similarity value and its corresponding index. We identify the most similar index between each attribute feature $A_i | i = 1, 2, \cdots, N$ and the global representations G, and then combine the attribute feature A_i and with its corresponding global representation G_k^t (where k is the matched index) to obtain the attribute-wise representations $AG^t = \{AG_i | i = 1, 2, \cdots, N\}$. This process is defined as:

$$AG_i^t = A_i^t + G_k^t. \tag{6}$$

We leverage the triplet loss to align attribute-wise representations with identity at the current step, assisting in global representations to distinguish similar identities.

$$L_{\text{Tri}}^l = max(d_p - d_n + m, 0). \tag{7}$$

where, d_p^l and d_n^l are the distances from positive samples and negative samples to anchor samples in attribute-wise representations, respectively. In this paper, attribute-wise representations that contain specific information of individuals assist global representations in distinguishing similar identities for maximizing intra-domain discrimination. Meanwhile, attribute-wise representations as a bridge across increasing datasets to minimize inter-domain gaps for better knowledge transfer.

E. Attribute-oriented Anti-Forgetting

We develop an attribute-oriented anti-forgetting (AF) strategy to explore attribute-wise representations that align the distributions of the old and new models, as shown in Fig. 2 (AF). The new model can adapt to new information but may forget old knowledge from the previous datasets, while the old model retains old knowledge. To preserve old knowledge, we leverage attribute-wise representations as a bridge to optimize both the old and new models by using samples from the memory buffer. This strategy achieves domain consistency and minimizes inter-domain gaps, alleviating the forgetting of old knowledge, and is calculated as follows:

$$L_{AF} = \frac{1}{B} \sum_{i=1}^{B} KL(AG_N^{t-1}/\tau || AG_N^t/\tau). \tag{8}$$

Where $KL(.||.)$ is a kullback-leibler divergence, and τ represents a hyper-parameter called temperature [41].

F. Knowledge Consolidation

Maximizing intra-domain discrimination and minimizing inter-domain gaps are in a contradictory relationship. Therefore, achieving a balance between them is crucial for improving the performance of LReID models. Thus, we propose a knowledge consolidation (KC) strategy that leverages global representations for knowledge transfer between old and new models. This includes alignment mechanism and logit-level distillation mechanism.

Maintaining distribution consistency between the old and new models for previous datasets can limit the model's ability to learn new knowledge. Therefore, we propose an alignment mechanism to explore global representations of knowledge transfer from the current dataset, as follows:

$$L_{AL} = \frac{1}{B} \sum_{i=1}^{B} KL(G^{t-1}/\tau || G^t/\tau). \tag{9}$$

We further introduce a logit-level distillation mechanism to enhance the extraction of identity information shared between the old and new models, further improving the model's knowledge consolidation ability. This is represented as follows:

$$L_{LD} = \frac{1}{B} \sum_{i=1}^{B} KL((\phi^{t-1}(G^{t-1}))_i/\tau || (\phi^t(G^t))_i/\tau). \tag{10}$$

The knowledge consolidation loss is defined as:

$$L_{KC} = L_{AL} + L_{LD}. \tag{11}$$

The total loss function is formulated as:

$$L = L_{CE} + L_{Tri}^g + L_{Tri}^l + L_{Ort} + L_{AF} + L_{KC}. \tag{12}$$

TABLE II
DATASET STATISTICS FOR BOTH SEEN AND UNSEEN DOMAINS. SINCE THE SELECTION PROCESS RESULTED IN 500 TRAIN IDS BEING SELECTED, THE ORIGINAL NUMBERS OF IDS ARE LISTED FOR COMPARISON. '-' DENOTES THAT THE DATASET IS NOT USED FOR TRAINING.

Type	Datasets	Scale	Train IDs	Test IDs
Seen	Market1501 [42]	large	500(751)	750
	CUHK-SYSU [43]	mid	500(942)	2900
	DukeMTMC [44]	large	500(702)	1110
	MSMT17_V2 [45]	large	500(1041)	3060
	CUHK03 [46]	mid	500(700)	700
Unseen	VIPeR [47]	small	–	316
	GRID [48]	small	–	126
	CUHK02 [49]	mid	–	239
	Occ_Duke [50]	large	–	1100
	Occ_REID [51]	mid	–	200
	PRID2011 [52]	small	–	649

IV. EXPERIMENTS

A. Experiments Setting

1) Datasets: To assess the performance of our method in anti-forgetting and generalization, we evaluate our method on a challenging benchmark consisting of Market1501 [42], CUHK-SYSU [43], DukeMTMC [44], MSMT17_V2 [45] and CUHK03 [46], referred to as the seen domain. Two representative training orders are set up following the protocol described in [22] for training and testing. Further, we employ six datasets including VIPeR [47], GRID [48], CUHK02 [49], Occ_Duke [50], Occ_REID [51], and PRID2011 [52], as the unseen domain. During evaluation, Unseen domain and test sets of the seen domain are combined into a single benchmark. Detailed statistics for these datasets can be shown in TABLE II.

2) Implementation Details: Our text encoder and image encoder are based on a pre-trained CLIP model, while the attribute decoder utilizes a transformer-based architecture [40]. All person images are resized to 256×128. We use Adam [54] for optimization and train each task for 60 epochs. The batch size is set to 128. The learning rate is initialized at 5×10^{-6} and is decreased by a factor of 0.1 every 20 epochs for each task. We employ mean average precision (mAP) and Rank-1 accuracy (R-1) to evaluate the LReID model on each dataset.

B. Comparison with SOTA Methods

We compare the proposed DCR with SOTA LReID to demonstrate the superiority of our method, including AKA [22], PTKP [9], PatchKD [16], KRKC [10], ConRFL [12], CODA [53], LSTKC [15], C2R [18], DKP [17]. Experimental results on training order-1 and order-2 are shown in TABLE III and TABLE IV, respectively.

1) Compared with LReID methods on the Seen Domain: In TABLE III and TABLE IV, our DCR significantly outperforms LReID methods, with an seen-avg incremental gain of 10.0% mAP/7.8% R-1, and 9.8% mAP/7.5% R-1 on training order-1 and order-2, respectively. Meanwhile, our DCR effectively alleviates catastrophic forgetting, achieving 6.9% mAp/1.1% R-1 and 5.4% mAP/2.2% R-1 improvement on the first dataset (Mrket1501 and DukeMTMC) with different training orders. Compared to CODA, our DCR significantly outperforms performance under the backbone of VIT-B/16. In contrast, our DCR achieves a trade-off between anti-forgetting and acquiring new information.

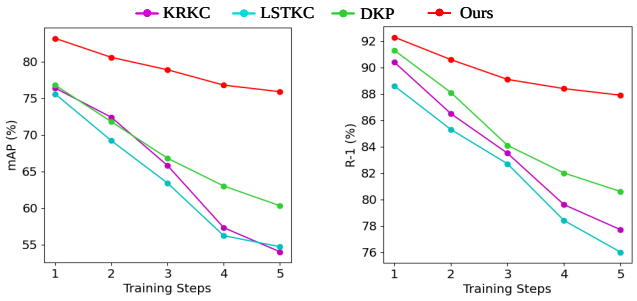


Fig. 3. Anti-forgetting curves. After each training step, we measure the metrics of Market1501 in the training order-1 to demonstrate the model's anti-forgetting performance.

2) Compared with LReID methods on the Unseen Domain: In Table III, compared to the KRKC method, our DCR model achieves significant improvements, with an average gain of 8.1% mAP / 7.5% R-1 on the unseen domain. Furthermore, when evaluated against the LSTKC methods, our DCR demonstrates superior performance over LReID methods, delivering an unseen-avg incremental improvement of 9.5% mAP/11.0% R-1, as shown in Table IV. We attribute these advancements to the complementary relationship between global and attribute-wise representations, which effectively enhances discrimination among similar identities. Additionally, our attribute-oriented anti-forgetting (AF) and knowledge consolidation (KC) strategies foster inter-domain consistency and facilitate seamless knowledge transfer. In summary, our DCR markedly strengthens generalization capabilities, outperforming existing approaches in terms of mAP and R-1 metrics.

3) Compared with Baseline: Due to the lack of CLIP-based comparison methods in LReID, we introduce a Baseline model that includes the CLIP model, an attribute-text generator, and a knowledge consolidation strategy. The Baseline outperforms other methods (such as AKA, PTKP, PatchKD, *etc.*) in mAP and R-1, benefiting from the powerful extraction capabilities of CLIP, as presented in TABLE III and TABLE IV. Compared to the Baseline, our DCR improves the Seen-Avg by 11.4% mAP/10.4% R-1 and by 9.8% mAP/10.2% R-1. These results demonstrate that our proposed domain consistency representation learning strategy achieves significant performance in balancing the maximization of intra-domain discrimination and the minimization of inter-domain gaps in LReID.

4) The Anti-forgetting Performance of Our Method: We conduct a forgetting measurement experiment in training order-1, as shown in Fig. 3. The Fig. 3 shows the metric measurements for the Market1501 dataset at different training steps. After training on the large-scale MSTMS17 dataset at training step 4, KRKC, LSTCC, and DKP exhibit significant attenuation in mAP and R-1. Because the comparison method limits the performance of the model in minimizing inter-

TABLE III
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON TRAINING ORDER-1. BOLD AND RED FONTS ARE OPTIMAL AND SUBOPTIMAL VALUES, RESPECTIVELY. TRAINING ORDER-1 IS MARKET1501→CUHK-SYSU→ DUKEMTMC→MSMT17_V2→CUHK03.

Method	Market1501		CUHK-SYSU		DukeMTMC		MSMT17_V2		CUHK03		Seen-Avg		Unseen-Avg	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
AKA [22]	58.1	77.4	72.5	74.8	28.7	45.2	6.1	16.2	38.7	40.4	40.8	50.8	42.0	39.8
PTKP [9]	64.4	82.8	79.8	81.9	45.6	63.4	10.4	25.9	42.5	42.9	48.5	59.4	51.2	49.1
PatchKD [16]	68.5	85.7	75.6	78.6	33.8	50.4	6.5	17.0	34.1	36.8	43.7	53.7	45.1	43.3
KRKC [10]	54.0	77.7	83.4	85.4	48.9	65.5	14.1	33.7	49.9	50.4	50.1	62.5	52.7	50.8
ConRFL [12]	59.2	78.3	82.1	84.3	45.6	61.8	12.6	30.4	51.7	53.8	50.2	61.7	-	-
CODA [53]	53.6	76.9	75.7	78.1	48.6	59.5	13.2	31.3	47.2	48.6	47.7	58.9	44.5	42.4
LSTKC [15]	54.7	76.0	81.1	83.4	49.4	66.2	20.0	43.2	44.7	46.5	50.0	63.1	51.3	48.9
C2R [18]	69.0	86.8	76.7	79.5	33.2	48.6	6.6	17.4	35.6	36.2	44.2	53.7	-	-
DKP [17]	60.3	80.6	83.6	85.4	51.6	68.4	19.7	41.8	43.6	44.2	51.8	64.1	49.9	46.4
Baseline	61.6	79.1	80.2	80.6	50.2	64.3	15.1	36.5	44.9	46.8	50.4	61.5	51.8	49.4
Ours	**75.9**	**87.9**	**87.3**	**88.5**	**60.1**	**71.9**	**25.3**	**50.1**	**60.5**	**61.3**	**61.8**	**71.9**	**60.8**	**58.3**

TABLE IV
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON TRAINING ORDER-2. BOLD AND RED FONTS ARE OPTIMAL AND SUBOPTIMAL VALUES, RESPECTIVELY. TRAINING ORDER-2 IS DUKEMTMC→MSMT17_V2→MARKET1501→ CUHK-SYSU→CUHK03.

Method	DukeMTMC		MSMT17_V2		Market1501		CUHK-SYSU		CUHK03		Seen-Avg		Unseen-Avg	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
AKA [22]	42.2	60.1	5.4	15.1	37.2	59.8	71.2	73.9	36.9	37.9	38.6	49.4	41.3	39.0
PTKP [9]	54.8	70.2	10.3	23.3	59.4	79.6	80.9	82.8	41.6	42.9	49.4	59.8	50.8	48.2
PatchKD [16]	58.3	74.1	6.4	17.4	43.2	67.4	74.5	76.9	33.7	34.8	43.2	54.1	44.8	43.3
KRKC [10]	50.6	65.6	13.6	27.4	56.2	77.4	83.5	85.9	46.7	46.6	50.1	61.0	52.1	47.7
ConRFL [12]	34.4	51.3	7.6	20.1	61.6	80.4	82.8	85.1	49.0	50.1	47.1	57.4	-	-
CODA [53]	38.7	56.6	11.6	24.5	54.3	75.1	76.2	75.8	42.3	41.7	44.6	54.7	45.0	42.9
LSTKC [15]	49.9	67.6	14.6	34.0	55.1	76.7	82.3	83.8	46.3	48.1	49.6	62.1	51.7	49.5
C2R [18]	59.7	75.0	7.3	19.2	42.4	66.5	76.0	77.8	37.8	39.3	44.7	55.6	-	-
DKP [17]	53.4	70.5	14.5	33.3	60.6	81.0	83.0	84.9	45.0	46.1	51.3	63.2	51.3	47.8
Baseline	53.8	69.1	14.1	29.8	59.8	80.4	78.4	78.5	45.3	44.9	50.3	60.5	52.2	49.9
Ours	**64.1**	**77.2**	**25.4**	**44.9**	**70.6**	**84.5**	**86.1**	**88.2**	**54.2**	**58.7**	**60.1**	**70.7**	**61.6**	**59.2**

domain gaps. Our method demonstrates a smoother decrease in mAP and Rank-1, which can effectively reduce inter-domain gaps to alleviate the catastrophic forgetting problem.

5) The effectiveness of minimizing inter-domain gaps: We visualize the feature distribution of PTKP, KRKC, Baseline, and our method across five datasets as shown in Fig. 4. The Baseline shows poor performance in bridging inter-domain gaps, as the lack of attribute-wise representations makes it challenging to reduce inter-domain gaps. The KRKC method effectively separates each domain, but it insufficiently distinguishes identity information within the domain, limiting the model's ability to prevent forgetting and enhance generalization. Compared to other methods, our DCR effectively distinguishes identity information within a domain and spreads identity information across multiple domains, which significantly improves the anti-forgetting and generalization ability of the model.

6) The effectiveness of maximizing intra-domain discrimination: We visualize the feature distribution of KRKC and our method. Fig. 5 shows that our DCR can significantly cluster images of the same identity more tightly (circle) and increase the distance between different identities (black bidirectional arrow). Compared to KRKC, our DCR improves intra-domain discrimination due to the complementary relationship between global and attribute-wise representations, which enables it to learn the subtle nuances of individuals.

7) Generalization Curves on the Unseen Domain: We analyze the average performance on the unseen domain during the training steps, as depicted in Fig. 6. Compared to other methods, our DCR achieves superior performance and exhibits faster performance growth across the training steps. Thus, our attribute-oriented anti-forgetting (AF) strategy effectively bridges inter-domain gaps and enhances the generalization ability of our model. In summary, our DCR explores global and attribute-wise representations to achieve a trade-off between maximizing intra-domain discrimination and minimizing inter-domain gaps.

C. Ablation Studies

1) The number of global and attribute-wise representations: Global and attribute-wise representations capture individual nuances in intra-domain and inter-domain consistency. We evaluate the suitability of multiple global and attribute-wise

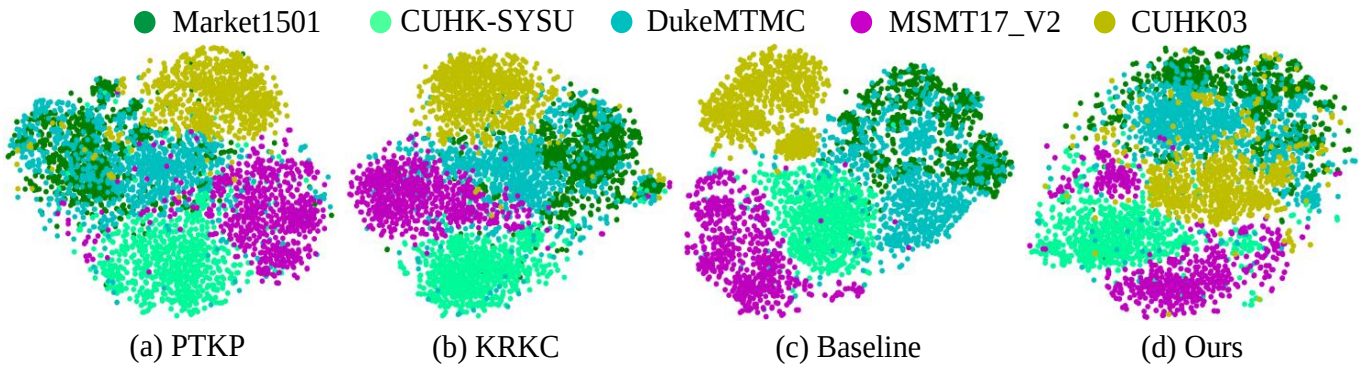


Fig. 4. t-SNE visualization of feature distribution on five datasets from the seen domain. Our DCR effectively improves the anti-forgetting and generalization ability, which significantly distinguishes identity information within a domain but also spreads identity information across multiple domains.

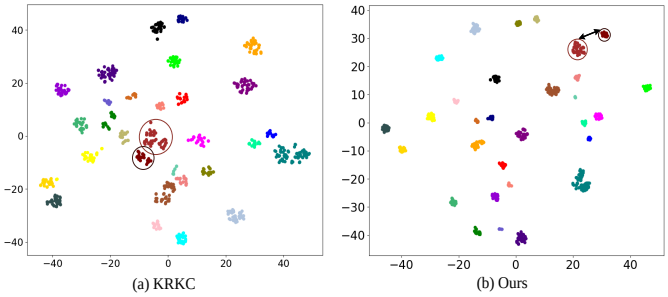


Fig. 5. Visualization of intra-domain discrimination on the Market1501 dataset. We randomly select 30 identities. Colors represent different identity information. Our DCR can cluster images of the same identity more tightly (circle) for minimizing inter-domian discrimination.

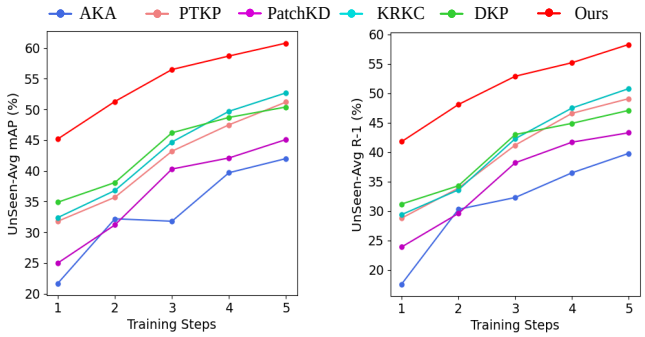


Fig. 6. Generalization curves. After each training step, the performance of the unseen domain is evaluated.

TABLE V
ABLATION STUDIES ON THE NUMBER OF GLOBAL AND ATTRIBUTE-WISE REPRESENTATIONS N ON TRAINING ORDER-1.

Number (N)	Seen_Avg		Unseen_Avg	
	mAP	R-1	mAP	R-1
2	60.2	68.7	59.4	56.5
3	**61.8**	**71.9**	**60.8**	**58.3**
4	61.2	71.6	60.3	57.5

representations as shown in TABLE V. We have observed that setting the number of global and attribute-wise representations N to 3 achieves the best performance for our method.

TABLE VI
ABLATION STUDIES OF DIFFERENT COMPONENTS ON TRAINING ORDER-1.

Type	Components				Seen_Avg		Unseen_Avg	
	PFM	ACN	AF	KC	mAP	R-1	mAP	R-1
(a)					50.4	61.5	51.8	49.4
(b)	√				51.7	62.1	52.5	50.3
(c)	√	√			56.9	68.3	57.6	55.4
(d)	√	√	√		58.7	69.2	58.5	56.8
(e)	√	√	√	√	**61.8**	**71.9**	**60.8**	**58.3**

2) Performance of Different Components: To evaluate the contribution of each component to our DCR model, we conduct ablation studies on both seen and unseen domains, as detailed in Table VI. Here, PFM denotes the Parallel Fusion Module, while ACN refers to the Attribute Compensation Network. AF and KC represent the attribute-oriented anti-forgetting and knowledge consolidation strategies, respectively. Comparing (a) and (b), we observe that PFM improves performance by +1.3% mAP / +0.6% R-1 on seen_avg and +0.7% mAP / +0.9% R-1 on unseen_avg under the baseline method. This demonstrates its effectiveness in fusing text and image information to generate diverse global representations. Comparing (a) and (c), integrating PFM and ACN yields significant gains: +6.5% mAP / +6.8% R-1 on seen_avg and +5.8% mAP / +6.0% R-1 on unseen_avg, for generating the diverse global representations. Comparing (c) and (e), when integrating AF and KC, the model achieves gains of +4.9% mAP / +3.6% R-1 on seen_avg and +3.2% mAP / +2.9% R-1 on unseen_avg. These strategies significantly promote inter-domain consistency and knowledge transfer. Our DRE method unifies PFM, ACN, AF, and KC into an end-to-end LReID model, striking a balance between maximizing intra-domain discrimination and minimizing inter-domain gaps. These combinations ensure strong generalization while mitigating catastrophic forgetting in our DCR.

3) Performance of attribute-text generator: To better understand whether each instance's text descriptions generated by the attribute-text generator (ATG) provide more fine-



Fig. 7. Visualization of retrieval results. The left column displays the query image under challenging conditions, such as occlusion, blur, side view, and back view. In the retrieved results, correctly matched pedestrians are marked with green boxes, while incorrect matches are highlighted in red box. Compared to the DKP method, our DCR retrieves more pedestrians with the same identity, demonstrating superior performance.

TABLE VII
ABLATION OF TRAINING WITH OR WITHOUT ATTRIBUTE-TEXT GENERATOR (ATG) ON TRAINING ORDER-1.

Method	Seen_Avg		Unseen_Avg	
	mAP	R-1	mAP	R-1
Training w/o ATG	60.1	70.5	59.3	56.5
Training w/ ATG	**61.8**	**71.9**	**60.8**	**58.3**

TABLE VIII
ABLATION STUDIES OF CONFIDENCE THRESHOLD ON TRAINING ORDER-1.

Confidence Threshold	Seen_Avg		Unseen_Avg	
	mAP	R-1	mAP	R-1
0.7	60.9	71.1	59.7	57.6
0.8	61.8	**71.9**	**60.8**	**58.3**
0.9	**62.0**	71.6	60.5	58.0

grained guidance for learning global representations, we train our model using the generic text descriptor "A photo of a person" (w/o ATG) for comparison. TABLE VII shows that the attribute-text generator obtains text descriptions to significantly improve overall performance. When using the specific text descriptors, the average decreases by 1.7% mAP/1.4% R-1 on the seen domain and by 1.5% mAP/1.8% R-1 on the unseen domain. ATG enhances the robustness of global representations for each instance, effectively mitigating the forgetting of old knowledge.

4) Analysis of confidence threshold: To evaluate the sensitivity of the confidence threshold in an attribute-text generation, we conduct an ablation study with threshold values of 0.7, 0.8, and 0.9, as shown in TABLE VIII. The results demonstrate that increasing the confidence threshold from 0.7 to 0.8 consistently enhances DCR performance across metrics. While a confidence threshold of 0.9 achieves the highest 62.0% mAP on seen_avg, other metrics do not reach their peak performance at this value. Overall, we find that setting the confidence threshold to 0.8 strikes the best balance, ensuring robust classification accuracy while maximizing model performance.

D. Visualization

To further validate the effectiveness of our DCR, we conduct a qualitative comparison with the DKP method, as shown in Fig. 7. The left column displays challenging query images, including cases with occlusion, blur, side views, and back views. In Fig. 7, correctly matched pedestrians are marked with green boxes, while incorrect matches are highlighted in red. Under occlusion conditions (first and second rows), our DCR accurately retrieves eight out of ten individuals with the same identity, despite the limited visual information available. For blurry queries (third row), the DKP method only correctly matches three pedestrians, whereas our DCR achieves perfect retrieval (10/10) even with the unclear query image. Additionally, in side-view and back-view scenarios (fourth and fifth rows), DCR successfully handles clothing color variations, outperforming DKP in matching the same identity. These results demonstrate that DCR achieves superior retrieval performance, consistently identifying more pedestrians with the same identity across diverse challenging conditions. This advantage stems from the collaborative design of global and attribute-wise representations, which balances intra-domain discrimination and inter-domain gaps. By enhancing discrimination between similar identities and adapting to complex scenarios, our DCR demonstrates significantly more robustness than existing approaches.

TABLE IX
EXPERIMENTAL RESULTS REGARDING TRAINING TIME AND INFERENCE TIME ON THE MARKET1501.

Method	Training time (h-min)	inference time (s)
AKA [22]	4h-50min	25s
PTKP [9]	16h-19min	20s
PatchKD [16]	3h-30min	103s
KRKC [12]	17h-31min	56s
DKP [17]	3h-50min	38s
Ours	19h-10min	65s

E. Analysis of Speed

We evaluate the computational efficiency and performance of various methods on the Market1501 dataset, which consists of a query set (750 IDs, 3,368 images) and a gallery set (751 IDs, 15,913 images). All experiments were conducted using two NVIDIA RTX A6000 GPUs for training and a single GPU for inference. As shown in Table IX, where time measurements are indicated in hours (h), minutes (min), and seconds (s), our DCR completes training in 19 hours and 10 minutes while achieving an inference time of 65 seconds, processing gallery images at a rate of 52 images per second. Although our DCR shows longer training and inference times compared to AKA, PatchKD, and DKP approaches, it delivers significantly superior re-identification performance, achieving an seen_avg incremental gain of 10.0% mAP/7.8% R-1, and 9.8% mAP/7.5% R-1, an seen_avg improvement 10.0% mAP/7.8% R-1, and 9.8% mAP/7.5% R-1 on training order-1 and order-2, as demonstrated in Tables III and IV. The increased computational complexity of DCR stems from its transformer-based CLIP architecture incorporating multiple class tokens and the additional attribute compensation network (ACN) that enhances discrimination of similar identities through local feature integration. While this design choice results in higher computational demands, it achieves an excellent balance between performance metrics and complexity. For future work, we plan to explore prompt tuning [53] and mixture-of-experts (MOE) approaches [55] to optimize the model's efficiency while maintaining its competitive performance advantages.

V. CONCLUSIONS

In this paper, we propose a domain consistency representation learning (DCR) model that explores global and attribute-wise representations to capture subtle nuances in intra-domain and inter-domain consistency, achieving a trade-off between maximizing intra-domain discrimination and minimizing inter-domain gaps. Specifically, global and attribute-wise representations serve as complementary information to distinguish similar identities within the domain. We also develop an attribute-oriented anti-forgetting (AF) strategy and a knowledge consolidation (KC) strategy to minimize inter-domain gaps and facilitate knowledge transfer, enhancing generalization capabilities. Extensive experiments demonstrate that our method outperforms state-of-the-art LReID methods.

REFERENCES

[1] A. Zhang, Y. Gao, Y. Niu, W. Liu, and Y. Zhou, "Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2021, pp. 598–607.

[2] Z. Pang, C. Wang, L. Zhao, Y. Liu, and G. Sharma, "Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 4, pp. 2706–2718, 2023.

[3] Z. Ran, X. Wei, W. Liu, and X. Lu, "Multi-scale aligned spatial-temporal interaction for video-based person re-identification," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 9, pp. 8536–8546, 2024.

[4] S. Chen, H. Da, D.-H. Wang, X.-Y. Zhang, Y. Yan, and S. Zhu, "Hasi: Hierarchical attention-aware spatio–temporal interaction for video-based person re-identification," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 6, pp. 4973–4988, 2024.

[5] Z. Yang, D. Wu, C. Wu, Z. Lin, J. Gu, and W. Wang, "A pedestrian is worth one prompt: Towards language guidance person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2024, pp. 17 343–17 353.

[6] H. Fan, X. Wang, Q. Wang, S. Fu, and Y. Tang, "Skip connection aggregation transformer for occluded person reidentification," *IEEE Trans. Ind. Inf*, vol. 20, no. 1, pp. 442–451, 2023.

[7] J. Tian, S. Zhou, B. Fan, and H. Zhang, "A novel image formation model for descattering," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 46, no. 12, pp. 8173–8190, 2024.

[8] S. Liu, H. Fan, Q. Wang, Z. Han, Y. Guan, and Y. Tang, "Wavelet–pixel domain progressive fusion network for underwater image enhancement," *Knowledge-Based Syst*, vol. 299, p. 112049, 2024.

[9] W. Ge, J. Du, A. Wu, Y. Xian, K. Yan, F. Huang, and W.-S. Zheng, "Lifelong person re-identification by pseudo task knowledge preservation," in *Proc. AAAI Conf. Artif. Intell*, vol. 36, no. 1, 2022, pp. 688–696.

[10] C. Yu, Y. Shi, Z. Liu, S. Gao, and J. Wang, "Lifelong person re-identification via knowledge refreshing and consolidation," in *Proc. AAAI Conf. Artif. Intell*, vol. 37, no. 3, 2023, pp. 3295–3303.

[11] N. Pu, Y. Liu, W. Chen, E. M. Bakker, and M. S. Lew, "Meta reconciliation normalization for lifelong person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 541–549.

[12] J. Huang, X. Yu, D. An, Y. Wei, X. Bai, J. Zheng, C. Wang, and J. Zhou, "Learning consistent region features for lifelong person re-identification," *Pattern Recognition*, vol. 144, p. 109837, 2023.

[13] G. Wu and S. Gong, "Generalising without forgetting for lifelong person re-identification," in *Proc. AAAI Conf. Artif. Intell*, vol. 35, no. 4, 2021, pp. 2889–2897.

[14] L. Zhang, G. Gao, and H. Zhang, "Spatial-temporal federated learning for lifelong person re-identification on distributed edges," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 35, no. 2, pp. 1884–1896, 2023.

[15] K. Xu, X. Zou, and J. Zhou, "Lstkc: Long short-term knowledge consolidation for lifelong person re-identification," in *Proc. AAAI Conf. Artif. Intell*, vol. 38, no. 14, 2024, pp. 16 202–16 210.

[16] Z. Sun and Y. Mu, "Patch-based knowledge distillation for lifelong person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 696–707.

[17] K. Xu, X. Zou, Y. Peng, and J. Zhou, "Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2024, pp. 16 604–16 613.

[18] Z. Cui, J. Zhou, X. Wang, M. Zhu, and Y. Peng, "Learning continual compatible representation for re-indexing free lifelong person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2024, pp. 16 614–16 623.

[19] S. Liu, H. Fan, Q. Wang, X. Chen, Z. Han, and Y. Tang, "Diverse representation embedding for lifelong person re-identification," *arXiv:2403.16003*, 2024.

[20] Y. Yan, H. Yu, Y. Wang, S. Song, W. Huang, and J. Jin, "Unified stability and plasticity for lifelong person re-identification in cloth-changing and cloth-consistent scenarios," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 10, pp. 9166–9180, 2024.

[21] Y. Xing, G. Xiao, M. S. Lew, and S. Wu, "Lifelong visible-infrared person re-identification via a tri-token transformer with a query-key mechanism," in *Proc. Int. Conf. Multimedia Retrieval*, 2024, pp. 988–997.

[22] N. Pu, W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew, "Lifelong person re-identification via adaptive knowledge accumulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2021, pp. 7901–7910.

[23] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[24] ——, "Conditional prompt learning for vision-language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2022, pp. 16 816–16 825.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[26] S. Yan, N. Dong, L. Zhang, and J. Tang, "Clip-driven fine-grained text-image person re-identification," *IEEE Trans. Image Process*, vol. 32, pp. 6032–6046, 2023.

[27] G. Han, M. Lin, Z. Li, H. Zhao, and S. Kwong, "Text-to-image person re-identification based on multimodal graph convolutional network," *IEEE Trans. Multimedia*, vol. 26, pp. 6025–6036, 2023.

[28] Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, and P. Hu, "Noisy-correspondence learning for text-to-image person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2024, pp. 27 197–27 206.

[29] Z. Shao, X. Zhang, C. Ding, J. Wang, and J. Wang, "Unified pre-training with pseudo texts for text-to-image person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2023, pp. 11 174–11 184.

[30] X. Wu, W. Ma, D. Guo, T. Zhou, S. Zhao, and Z. Cai, "Text-based occluded person re-identification via multi-granularity contrastive consistency learning," in *Proc. AAAI Conf. Artif. Intell*, vol. 38, no. 6, 2024, pp. 6162–6170.

[31] S. Li, L. Sun, and Q. Li, "Clip-reid: exploiting vision-language model for image re-identification without concrete text labels," in *Proc. AAAI Conf. Artif. Intell*, vol. 37, no. 1, 2023, pp. 1405–1413.

[32] Y. Du, Z. Zhao, and F. Su, "Yyds: Visible-infrared person re-identification with coarse descriptions," *arXiv: 2403.04183*, 2024.

[33] J. Jia, H. Huang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting," *arXiv:2107.03576*, 2021.

[34] H. Sun, H. Zhao, W. Zhang, L. Xu, and H. Guan, "Adaptive multi-task learning for multi-par in real-world," *IEEE J. Radio Freq. Identif*, 2024.

[35] Y. Zhou and X. Zeng, "Towards comprehensive understanding of pedestrians for autonomous driving: Efficient multi-task-learning-based pedestrian detection, tracking and attribute recognition," *Rob. Auton. Syst*, vol. 171, p. 104580, 2024.

[36] X. Fan, Y. Zhang, Y. Lu, and H. Wang, "Parformer: Transformer-based multi-task network for pedestrian attribute recognition," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 1, pp. 411–423, 2023.

[37] J. Jia, X. Chen, and K. Huang, "Spatial and semantic consistency regularizations for pedestrian attribute recognition," in *Proc. Int. Conf. Comput. Vis*, 2021, pp. 962–971.

[38] J. Jia, N. Gao, F. He, X. Chen, and K. Huang, "Learning disentangled attribute representations for robust pedestrian attribute recognition," in *Proc. AAAI Conf. Artif. Intell*, vol. 36, no. 1, 2022, pp. 1069–1077.

[39] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proc. Int. Conf. Comput. Vis*, 2017, pp. 350–359.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Adv. Neural Inform. Process. Syst*, vol. 30, 2017.

[41] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015.

[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. Int. Conf. Comput. Vis*, 2015, pp. 1116–1124.

[43] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," *arXiv:1604.01850*, 2016.

[44] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis*, 2016, pp. 17–35.

[45] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2018, pp. 79–88.

[46] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2014, pp. 152–159.

[47] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis*, 2008, pp. 262–275.

[48] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *Int. J. Comput. Vision*, vol. 90, pp. 106–129, 2010.

[49] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2013, pp. 3594–3601.

[50] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. Int. Conf. Comput. Vis*, 2019, pp. 542–551.

[51] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proc. Int. Conf. Multimedia and Expo*, 2018, pp. 1–6.

[52] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scandinavian Conference on Image Analysis*, 2011, pp. 91–102.

[53] J. S. Smith, L. Karlinsky, V. Gutta, P. Cascante-Bonilla, D. Kim, A. Arbelle, R. Panda, R. Feris, and Z. Kira, "Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2023, pp. 11 909–11 919.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[55] J. Yu, Y. Zhuge, L. Zhang, P. Hu, D. Wang, H. Lu, and Y. He, "Boosting continual learning of vision-language models via mixture-of-experts adapters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2024, pp. 23 219–23 230.



Shiben Liu received his B.E. and M.S. degrees in Electronic Information Engineering, and Communication and Information Systems from Liaoning University of Engineering and Technology, Fuxin, China, in 2019 and 2022, respectively. He is currently pursuing a Ph.D. degree in mode recognition and intelligent systems from the Chinese Academy of Sciences University, Beijing, China. His current research focuses on deep learning, lifelong learning, person re-identification, image restoration and analysis. He has served as the Reviewer for the international journals such as the TCSVT, TMM, RAL, Neurocomputing, et al.



Huijie Fan (Member, IEEE) received the B.E. degree in automation from the University of Science and Technology of Science and Technology of China, China, in 2007, and the Ph.D. degree in mode recognition and intelligent systems from the Chinese Academy of Sciences University, Beijing, China, in 2014. She is currently a Research Scientist with the Institute of Shenyang Automation of the Chinese Academy of Sciences. Her research interests include deep learning on image processing and medical image processing and applications.



Qiang Wang received the B.E. and M.S. degrees in school of Computer Science and Technology from Shandong Jianzhu University and Tianjin Normal University, P.R.China,in 2004 and 2008, respectively, the Ph.D degree in State Key Laboratory of Robotics, Shenyang Institute of Automation, University of Chinese Academy of Sciences, Beijing, China in 2020. He is an Associate Professor in the Key Laboratory of Manufacturing Industrial Integrated in Shenyang University. He has some top-tier journal papers accepted at TIP, TMM, TCSVT, IoT-J and Pattern Recognition et al. His current research focuses on deep learning, multi-task learning, image restoration and analysis.



Weihong Ren received the B.E. degree in Automation and Electronic Engineering from Qingdao University of Science and Technology, Qingdao, China, in 2013. In 2020, he received Ph.D degree from the City University of Hong Kong, Hongkong, China, and the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China. He is an Assistant Professor at the School of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen, China. He also has some top-tier journals and conference papers accepted at IEEE TIP, IEEE TMM, CVPR, AAAI et al. His current research interests include object tracking, action recognition and deep learning.



Yandong Tang (Member, IEEE) received the B.S. and M.S. degree in the calculation of mathematics from Shandong University, China, in 1984 and 1987. From 1987 to 1996, he worked at the Institute of Computing Technology, Shenyang, Chinese Academy of Sciences. From 1996 to 1998, he was engaged in research and development at Stuttgart University and Potsdam University in Germany. He received a Ph.D. degree in Engineering Mathematics from the Research Center (ZETEM) of Bremen University, Germany, in 2002. From 2002 to 2004, he worked at the Institute of Industrial Technology and Work Science (BIBA) at Bremen University of Germany. He is currently a Research Scientist with the Institute of Shenyang Automation of the Chinese Academy of Sciences. His research interests include image processing, mode recognition and robot vision.



Yang Cong (Senior Member, IEEE) received the B.Sc. degree from Northeast University in 2004 and the Ph.D. degree from the State Key Laboratory of Robotics, Chinese Academy of Sciences, in 2009. From 2009 to 2011, he was a Research Fellow with the National University of Singapore (NUS) and Nanyang Technological University (NTU). He was a Visiting Scholar with the University of Rochester. He was the professor until 2023 with Shenyang Institute of Automation, Chinese Academy of Sciences. He is currently the full professor with South China University of Technology. He has authored over 80 technical articles. His current research interests include robot, computer vision, machine learning, multimedia, medical imaging and data mining. He has served on the editorial board of the several joural papers. He was a senior member of IEEE since 2015.