

# Distribution-Aware Knowledge Aligning and Prototyping for Non-Exemplar Lifelong Person Re-Identification

Jiahuan Zhou<sup>1b</sup>, Kunlun Xu<sup>1b</sup>, Fan Zhuo<sup>1b</sup>, Xu Zou<sup>1b</sup>, *Member, IEEE*, and Yuxin Peng<sup>1b</sup>, *Senior Member, IEEE*

**Abstract**—Lifelong person re-identification (LReID) suffers from the catastrophic forgetting problem when learning from non-stationary data streams. Existing exemplar-based and knowledge distillation-based LReID methods encounter data privacy and limited acquisition capacity, respectively. In this paper, we introduce the prototype, which is under-investigated in LReID, to better balance knowledge retention and acquisition. Previous prototype-based works primarily focused on the classification task, where prototypes were modeled as discrete points or statistical distributions. However, they either discarded the distribution information or omitted instance-level diversity, which are crucial fine-grained clues for LReID. Furthermore, the domain shifts between data sources result in a feature gap between the new and old data, which restricts the utilization of the fine-grained information in prototypes. To address these challenges, we propose Distribution-aware Knowledge Aligning and Prototyping (DKP++), a novel framework for modeling and leveraging prototypes in LReID. First, an Instance-level Distribution Modeling network is introduced to capture the local diversity of each instance. Next, a Distribution-oriented Prototype Generation algorithm transforms the instance-level diversity into identity-level distributions which are stored as prototypes. Then, a Prototype-based Knowledge Transfer module distills the knowledge within the prototypes to the new model. To mitigate the impact of domain shifts during knowledge transfer, we introduce a privacy-friendly Distribution Aligning module that transforms new input data to fit the historical distribution, which is incorporated with feature-level alignment constraints to enhance the coherence between new and old knowledge, effectively improving historical prototype utilization. Extensive experiments demonstrate that our method achieves a superior balance between plasticity and stability, outperforming state-of-the-art LReID methods by a large margin.

**Index Terms**—Lifelong learning, person re-identification, distribution modeling, knowledge distillation.

## I. INTRODUCTION

AS A conventional task in computer vision, person re-identification (ReID) [5], [6] has achieved remarkable performance in static scenarios where the training data are provided all at once for learning. However, the real-world deployment of these ReID models in dynamic scenarios, particularly in the context of extensive streaming data generated by surveillance systems, exposes a significant performance limitation [7], [8]. Therefore, recent efforts have shifted towards addressing a more challenging and practical lifelong person re-identification (LReID) problem [7], [9], where the training data from different scenarios comes continually and the trained model can not only acquire new information but also preserve previously acquired knowledge.

As well-known in existing lifelong learning tasks [10], how to alleviate catastrophic forgetting is also the bottleneck for LReID. Note that this catastrophic forgetting phenomenon becomes even more critical in LReID due to the large intra-person divergences and small inter-person variations. Specifically, the data distribution for the same person in LReID can exhibit significant variation, stemming from domain shifts across time and environments. Moreover, different persons may exhibit subtle nuances of individual information, which results in severe distribution overlapping, causing the forgetting of valuable discrimination knowledge for each individual.

To tackle this crucial issue, most existing LReID methods leverage additional memory to reserve past exemplars for reusing when learning new data [7], [8], [11], but inevitably raise severe privacy concerns and computational costs [12]. As an alternative, various non-exemplar methods [9], [12], [13], [14], [15] have been proposed to perform knowledge distillation by imposing output consistency constraints when encountering new datasets. However, such strict constraints would seriously limit the plasticity of the ReID model and lead to insufficient acquisition of the new domain data.

Recently, a few works have aimed to explore prototypes to mitigate catastrophic forgetting in other lifelong learning tasks, e.g., class incremental learning (CIL) [10]. These methods [2], [3], [10], [16] usually calculate or learn the feature center of a class as the prototype, as shown in Fig. 1(a). Nevertheless, solely

Received 2 October 2024; revised 28 May 2025; accepted 26 July 2025. Date of publication 7 August 2025; date of current version 5 November 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62376011, Grant 62525201, Grant 62132001, and Grant 62432001, in part by Beijing Natural Science Foundation under Grant L247006, and in part by the National Key R&D Program of China under Grant 2024YFA1410000. Recommended for acceptance by R. Cucchiara. (*Corresponding author: Yuxin Peng.*)

Jiahuan Zhou, Kunlun Xu, and Yuxin Peng are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China (e-mail: jiahuanzhou@pku.edu.cn; xkl@stu.pku.edu.cn; pengyuxin@pku.edu.cn).

Fan Zhuo is with the School of Cyber Engineering, Xidian University, Xi'an 710126, China (e-mail: fanzhuo@stu.xidian.edu.cn).

Xu Zou is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: zoux@hust.edu.cn).

The code is available at [https://github.com/zhouliahuan1991/TPAMI-DKP\\_Plus\\_Plus](https://github.com/zhouliahuan1991/TPAMI-DKP_Plus_Plus)

Digital Object Identifier 10.1109/TPAMI.2025.3597023

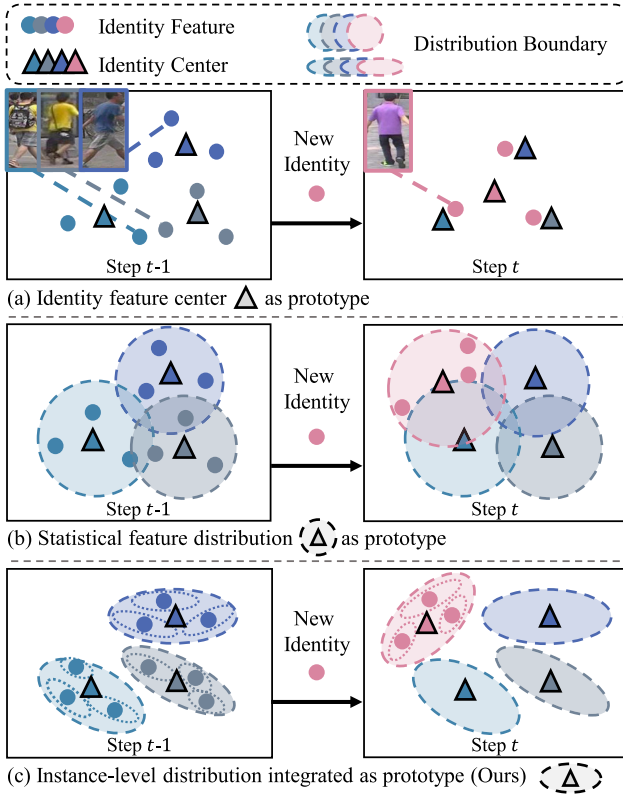


Fig. 1. The comparison of different prototype modeling solutions across training steps. (a) Various methods [1], [2], [3] treat the prototypes as category-aware feature center points but ignore the important distribution information. (b) Some approaches [4] calculate the prototype as a statistical feature distribution but suffer from inaccurate distribution estimation. (c) Our method models the instance-level distribution of each sample, from which the obtained distribution-aware prototypes are more informative and discriminative.

using a feature point as the prototype ignores the intra-class diversity, resulting in serious forgetting due to insufficient information retention of historical data [4]. Thus, several methods [4], [17], [18] consider calculating a statistical feature distribution of all sample features from the same class to enhance the representation ability of prototypes. However, as illustrated in Fig. 1(b), they simply treat all samples equally and neglect the unique characteristics of each individual. For example, the number of person samples captured from different viewpoints can vary significantly, with a sample from a rare-shot view providing more distributional information than a sample from a multi-shot view. Therefore, the prototypes obtained by existing methods inevitably drift from the real data distribution, misleading subsequent learning by the conveyed inaccurate knowledge, especially for LReID that aims to achieve fine-grained matching.

Moreover, as illustrated in Fig. 2(a), there often exists a natural domain shift between data collected from different periods or environments. This discrepancy results in a considerable separation between the feature distributions of the two sources. Such divergence weakens the coherence between historical and newly acquired feature representations, thereby reducing the effectiveness of historical prototypes in guiding the learning of new data. As a result, the model's ability to mitigate forgetting is substantially compromised, and the potential of historical

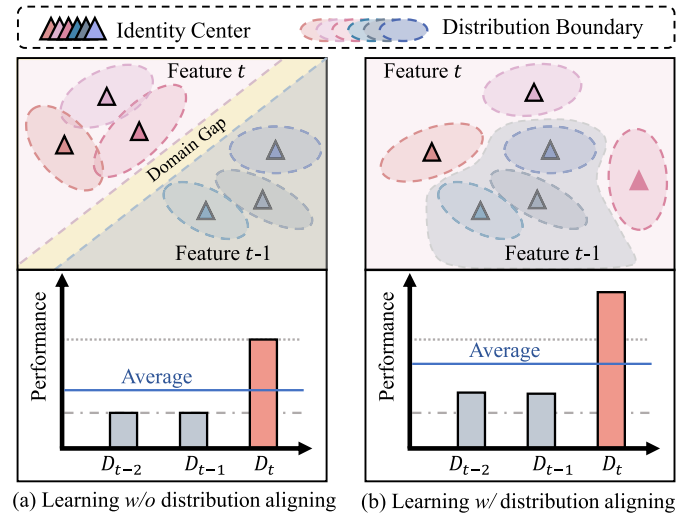


Fig. 2. (a) Due to the domain gap between the learning steps, most of the new features exhibit significant drift from the old data features, leading to limited utilization of the historical prototypes. (b) By introducing distribution aligning, the correlation between new features and historical prototypes is significantly improved. Therefore, the historical prototypes can be better leveraged to guide the new distribution modeling and mitigate catastrophic forgetting.

knowledge to support new data modeling is not adequately realized.

To settle the above issues, we propose a novel non-exemplar LReID method named Distribution-aware Knowledge Aligning and Prototyping (DKP++). The main idea comprises two aspects: First, model the instance-level distribution of each sample to generate more informative identity-specific prototypes as shown in Fig. 1(c). Second, adopt distribution alignment to reduce the gaps between the new and old features, thereby enhancing the utilization of the prototypes to improve anti-forgetting and discriminative new feature mining, as shown in Fig. 2(b). To achieve these, an Instance-level Distribution Modeling (IDM) network is designed to estimate the distribution for each input instance under the guidance of the proposed distribution-aware losses. Besides, to incorporate the prototypes with comprehensive instance-specific knowledge, we propose a Distribution-oriented Prototype Generation (DPG) strategy that transforms the instance-level distributions into a multivariate Gaussian distribution which is registered as the distribution-aware prototype. To effectively utilize the preserved knowledge in prototypes, a Prototype-based Knowledge Transfer (PKT) module is explored to guide the new model learning via enhancing the discriminability of new identity features with the aid of historical distributions. Furthermore, to overcome the impact of domain shift on knowledge transfer, we introduce a dual-side alignment scheme. On one hand, a Distribution Aligning Module (DAM) is developed to enhance the coherence between new and historical knowledge by transforming new images to match the historical image distribution through a lightweight and privacy-friendly distribution transfer network. On the other hand, by jointly leveraging both the original and aligned data with feature-level alignment constraints, the proposed approach not only improves the acquisition of new knowledge but also significantly mitigates catastrophic forgetting. Experimental results

demonstrate that our method outperforms existing approaches by substantial margins.

In summary, our contributions are four-fold: (1) A novel non-exemplar LReID method DKP++, is proposed that introduces the prototype into the LReID task. (2) A distribution-aware knowledge prototyping mechanism is developed where the instance-specific distributions are mined and transformed into prototypes to enhance historical knowledge retention. (3) A novel dual-side alignment scheme is introduced, which introduces privacy-friendly input-level alignment and joint feature-level alignment simultaneously to overcome the domain drift issue, significantly enhancing new knowledge mining and historical knowledge retention. (4) We provide a thorough theoretical analysis and comprehensive experiments to validate the superiority of our proposed LReID model against the state-of-the-art LReID approaches.

This manuscript is an extension of our previous conference paper [19]. In this study, we have made extensive extensions including (1) To mitigate the interference of domain shifts during historical prototype utilization, a privacy-friendly input-level alignment mechanism is developed to bridge the gap between the new and old domains. (2) To fully leverage the new and aligned data, feature-level constraints are introduced to form a dual-side alignment scheme, enhancing the model's capacity to acquire new knowledge and retain historical knowledge. (3) Comprehensive theoretical and experimental analyses are provided in Section IV to validate the effectiveness of our algorithm compared to existing works. (4) Five additional LReID dataset orders beyond the two used in the original study are introduced, demonstrating the robustness of our method under varied domain shift scenarios. (5) More ablation studies and detailed analyses, including learning trends, ablation studies on module effectiveness, adaptability to the vision foundation model, and visualization results, are presented in Section V.

## II. RELATED WORK

### A. Lifelong Person Re-Identification

Person Re-Identification [5], [20], [21], [22] aims to retrieve a specific individual from a given set of person images. Over the past decade, most ReID research has primarily concentrated on static settings, where the test environments and camera configurations are identical to those in the training phase [23], [24], [25]. However, in practical applications, new devices and changing environments frequently emerge, resulting in significant performance degradation of these models [26], [27], [28], [29]. To address this limitation, a more challenging and practical task, lifelong person re-identification [9], [30], [31], has been introduced, garnering increasing attention in recent years. The key challenge in LReID [10], [32] is the catastrophic forgetting problem, wherein the model's performance on previously learned domains degrades significantly when it is trained on new domains. To tackle this, existing LReID approaches can be broadly categorized into two types: exemplar-based and knowledge distillation-based methods.

Exemplar-based methods [7], [8], [11], [33] aim to mitigate catastrophic forgetting by storing images from previously

encountered domains and replaying them during the learning of new data. For example, GwFreID [7] utilized a herding selection strategy [34] to choose two exemplars per identity, while PTKP [8] and KRKC [11] randomly selected 250 identities, preserving two exemplars for each. Additionally, CLUDA-ReID [33] maintained a memory buffer containing 512 exemplars. Although storing historical exemplars provides an intuitive solution for mitigating forgetting and has demonstrated strong performance [35], the practical implementation of this approach in real-world applications is often constrained by privacy concerns and storage limitations [12], [36]. In light of these concerns, we adopt the non-exemplar LReID setting in this work.

Knowledge distillation-based methods [9], [12], [13], [14], [31] attempt to preserve historical knowledge by enforcing output consistency between the old and new models. For example, AKA [9] employed logit distillation to restrict the output distributions of the new model. PatchKD [13] introduced adaptively selected patches and developed a patch distillation module to transfer local knowledge from the old model to the new one. Building on PatchKD, C<sup>2</sup>R [31] further introduced balanced compatible distillation loss and balanced anti-forgetting distillation loss. LSTKC [14] proposed an affinity-based distillation loss to maintain the learned inter-instance relational structure. Though promising anti-forgetting capacity has been exhibited, their ability to acquire new data is limited due to the strict constraints that reduce model plasticity.

In this paper, to overcome the drawbacks of the exemplar-based and knowledge distillation-based methods, we propose a non-exemplar approach that introduces prototypes to settle the LReID problem, where the instance-level fine-grained information mining and advanced prototype utilization mechanisms are developed to facilitate knowledge accumulation.

### B. Prototype-Based Class Incremental Learning

One of the tasks closely related to LReID is class incremental learning (CIL), as each identity can be considered a category. Recently, various prototype-based CIL methods have been introduced to continually learn new classes without preserving historical exemplars [4], [16], [37], [38]. Some approaches treat the prototype of each class as a learnable embedding vector [39], [40], [41], while others take the mean feature of all samples from the same class as the prototype [3], [4], [37]. During the learning of new classes, these prototypes are used to represent the knowledge of previous classes [37]. However, relying on a single feature center point often results in a loss of informative data distribution. To address this, recent methods [4], [18], [42] have proposed calculating both the features' mean and variance to capture a more comprehensive representation of a class's distribution. However, these approaches assumed that all samples contributed equally to the estimated distribution, ignoring intra-class sample diversity. Consequently, the resulting prototypes suffered from distribution drift, which can mislead subsequent learning steps. Among these works, Fecam [18] introduced an anisotropic Mahalanobis matrix to describe class distributions and adopted a Bayesian classifier to accommodate continual data streams. However, since the feature extractor in Fecam is kept frozen, the

model fails to exploit instance-level distribution cues adequately, resulting in limited expressiveness of the estimated distributions. Furthermore, the frozen representation hinders the adaptation to new domains, as the testing data cannot benefit from improved feature representations. In this paper, we propose to integrate distribution modeling into prototype generation, allowing the storage of richer historical information and providing more precise guidance to the model during the subsequent data learning.

### C. Distribution Learning

Distribution plays a critical role in describing the probabilistic knowledge of data [43], [44]. Existing distribution learning methods primarily focus on modeling data uncertainty to handle out-of-distribution data [45], [46], [47], [48] or enhance matching accuracy [49]. For instance, [44] estimated label uncertainty in domain adaptive semantic segmentation to rectify noisy labels. Similarly, [45] leveraged distribution learning to mitigate the negative effects of label noise and outliers. Furthermore, [48] utilized distribution learning to enhance adversarial sample generation, while [49] introduced probabilistic regression to improve object tracking performance. Besides, [50] learned probability densities of face images to enhance matching accuracy. These works show that distribution modeling introduces abundant information for subsequent operations. Inspired by this, we propose to integrate the instance-level distributions into prototypes to enhance historical information storage. Then the historical prototypes are utilized to guide new knowledge learning and mitigate the catastrophic forgetting problem.

### D. Style Transfer in Person Re-Identification

Image-level style transfer has been actively explored in ReID, primarily for intra-domain learning [24], [51] and domain adaptation [21], [26], [52], [53]. Intra-domain methods aim to augment the training data via inter-camera style transfer [24] or learn a unity style transfer model to smooth the style disparities [51]. Domain adaptation approaches, on the other hand, focus on transferring styles between labeled and unlabeled domains to facilitate learning in the target (unlabeled) domain [52]. However, these methods typically assume concurrent access to data from both domains, which does not hold in the LReID setting. Moreover, both intra-domain learning and domain adaptation methods generally rely on deep generative networks to perform style transfer. Recent studies on model inversion attacks [54], [55] highlight the privacy risks associated with such generative models, raising concerns about their applicability in LReID scenarios. Some recent works have investigated feature space alignment [8], [56], which assumes that domain gaps are primarily encoded in the feature statistics. They usually perform cross-domain transfer by aligning feature-level statistical properties between domains [8]. However, since the model parameters are continually updated during training on new domains, the actual feature statistics of past domains inevitably drift from their recorded values [57]. As a result, feature alignment approaches also suffer from severe catastrophic forgetting in the lifelong learning setting. In this paper, we propose a lightweight style

transfer network that generates old-style data in a privacy-preserving manner, effectively mitigating the impact of feature drift during lifelong learning.

## III. THE PROPOSED DKP METHOD

### A. Problem Formulation and Notations

In non-exemplar LReID, a stream of  $T$  training datasets  $\mathcal{D} = \{D_t\}_{t=1}^T$  collected from different domains are provided step by step. At the  $t$ th training step, the images of the previous  $t-1$  domains are unavailable. Given a dataset  $D_t$  with  $N_t$  identities, our method obtains a prototype set  $\mathcal{P}_t = \{\mathbf{p}_i\}_{i=1}^{N_t}$  for all identities. For each instance, our method learns a multivariate Gaussian distribution  $\mathcal{N}(\mathbf{c}, \mathbf{v}^2)$ , where  $\mathbf{c} \in \mathbb{R}^d$  is the feature center vector of dimension  $d$ , and  $\mathbf{v}^2 \in \mathbb{R}^d$  is the diagonal element vector derived from the diagonal covariance matrix  $\Sigma \in \mathbb{R}^{d \times d}$ . The LReID model learned at the  $t$ th step is denoted  $\mathbf{M}_t$ . Besides, we also train a Distribution Transfer Network (DTNet)  $\Phi_t$  aiming to transfer the arbitrary input image  $x_i$  into the distribution of domain  $t$ , where the generated image is denoted  $x'_i$ . To enable  $\Phi_t$  training, the input image  $x_i$  is applied distribution augmentation to generate  $\tilde{x}_i$ .

### B. The Overall Pipeline of DKP++

At the  $t$ th learning step, the overall pipeline of our proposed DKP++ method, as depicted in Fig. 3(a) and (b), contains two training stages. The first stage aims to train a new LReID model, where the old distribution aligning is conducted, followed by an instance-level distribution modeling module (IDM) and a prototype-based knowledge transfer module (PTK). The second stage focuses on preparing the new prototypes and a new distribution transfer model for subsequent training steps.

### C. Stage1: LReID Model Training

1) *Distribution Aligning Module and LReID Model*: Firstly, we propose to align the distribution of the new data with the old data to overcome the distribution gap between the training steps. Specifically, as shown in Fig. 3(a), given a set of  $n$  input images  $\{x_i\}_{i=1}^n$  from  $D_t$ , a DTNet model  $\Phi_{t-1}$  trained in the previous training step is exploited, which transfers each input image  $x_i$  into  $x'_i$ , an image of the old distribution. The training process of DTNet is introduced in Section III-D.

The LReID model of DKP++ is built based on a CNN network, as shown in Fig. 3(a). At  $t$ th training step, the new model  $\mathbf{M}_t$  predicts the distribution centers of the input images  $x_i$  and  $x'_i$ , which are denoted as  $\mathcal{C}_t = \{\mathbf{c}_i\}_{i=1}^n$  and  $\mathcal{C}'_t = \{\mathbf{c}'_i\}_{i=1}^n$ , respectively. Besides, a variance predictor  $\Psi_t$  which comprises a convolution layer, followed by a pooling layer and a linear layer, is adopted to estimate the distribution variances of the instances  $\mathcal{V}_t = \{\mathbf{v}_i^2\}_{i=1}^n$  and  $\mathcal{V}'_t = \{(\mathbf{v}'_i)^2\}_{i=1}^n$ . When training the new LReID model  $\mathbf{M}_t$ , we maintain an old LReID model  $\mathbf{M}_{t-1}$  as the teacher model following previous works. The distribution-aligned data  $x'_i$  is fed to  $\mathbf{M}_{t-1}$  to obtain the predicted instance center  $\mathcal{C}_{t-1} = \{\mathbf{c}_i^o\}_{i=1}^n$ .

In addition to input-level distribution alignment, we further introduce the model output alignment to enhance the historical

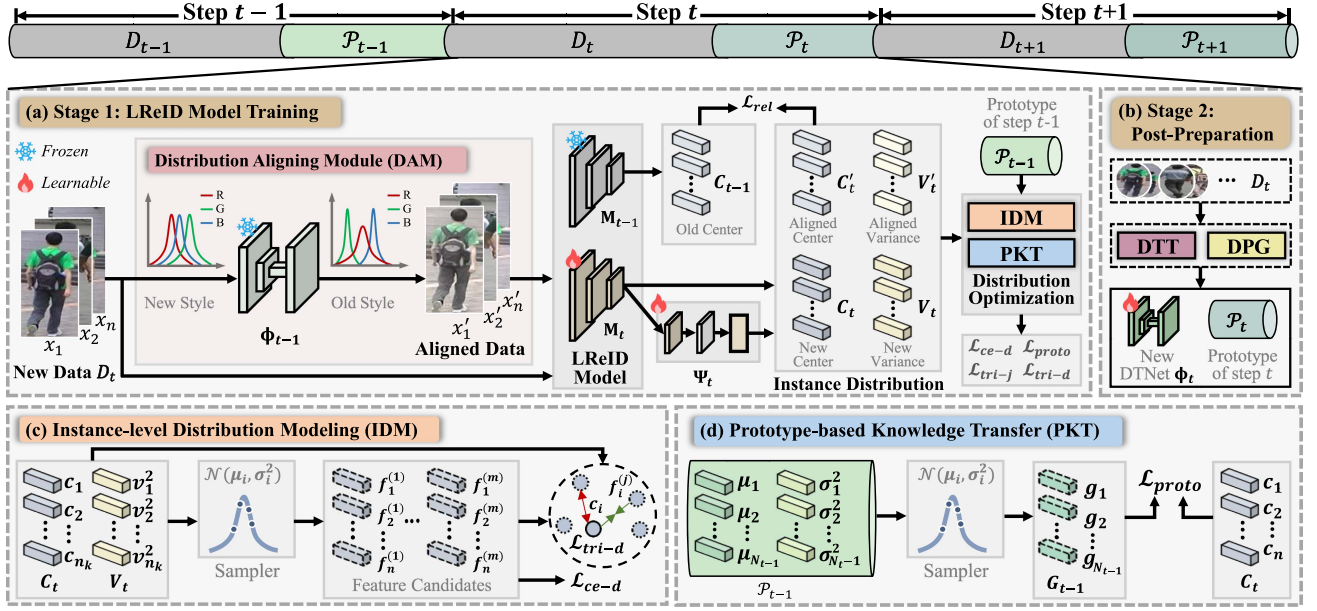


Fig. 3. The pipeline of our proposed Distribution-aware Knowledge Aligning and Prototyping (DKP++) approach. DKP++ contains two stages at each training step. In the first stage (a), the input images are transferred into a historical distribution by DAM. Then, the instance-specific feature centers and variances are modeled under the supervision of IDM and PKT modules. The second stage (b) is for post-preparation. The new data is further utilized to generate a new distribution transfer network (DTNet) and new prototypes  $P_t$  utilizing a Distribution Transfer Training module (DTT) and a Distribution-oriented Prototype Generation strategy (DPG). The IDM module (c) aims to guide the distribution learning of each instance. The PKT module (d) utilizes the previous step prototypes  $P_{t-1}$  to mitigate forgetting.

knowledge retention capacity. Specifically, a relation distribution alignment loss [14] is adopted:

$$\mathcal{L}_{rel} = \mathcal{L}_{KL}(\rho(C_{t-1}C_{t-1}^\top/\tau) \parallel \rho(C_tC_t^\top/\tau)), \quad (1)$$

where  $C_{t-1} \in \mathbb{R}^{n \times d}$  and  $C_t \in \mathbb{R}^{n \times d}$  are the feature matrix of instance centers predicted by the old and new models respectively. The softmax function  $\rho$  is applied row-wisely and  $\tau$  is the temperature parameter [58] to scale the matrix values.  $\mathcal{L}_{KL}$  is the Kullback Leibler divergence [59].

2) *Instance-level Distribution Modeling:* In this section, we aim to guide the model to learn the instance-level distribution, which contains fine-grained information by exploiting new data  $D^t$  and aligned data  $D^t$ .

*Intra-distribution Modeling:* To motivate the model learning, we adopt the Monte Carlo strategy [60] where the instance features are sampled from the predicted distributions.

Take the process of the new data as an example, as illustrated in Fig. 3(c), once the instance centers  $C_t$  and variances  $V_t$  are predicted, we adopt a Gaussian sampler parameterized as  $\mathcal{N}(c_i, v_i^2)$  to sample  $m$  feature candidates  $\{f_i^{(j)}\}_{j=1}^m$  for each instance, where each  $f_i^{(j)}$  is assigned the same identity label as  $c_i$ . To ensure identity consistency between  $\{f_i^{(j)}\}_{j=1}^m$  and  $c_i$ , a distribution-aware cross-entropy loss  $\mathcal{L}_{ce-d}$  is calculated as below:

$$\mathcal{L}_{ce-d} = \frac{1}{m+1} \left[ y_i \log \rho(W_t c_i) + \sum_{j=1}^m y_i \log \rho(W_t f_i^{(j)}) \right], \quad (2)$$

where  $y_i$  is the identity label of image  $x_i$ ,  $\rho$  represents the softmax function, and  $W_t$  denotes the linear projection parameter to map the features to logits.

Moreover, we further extend the widely used triplet loss in existing LReID models [11], [13] to a distribution-oriented version, denoted as  $\mathcal{L}_{tri-d}$ . Given  $n$  images  $\{x_i\}_{i=1}^n$  and their sampling set  $\mathcal{F}_s = \{(f_i^{(j)}, y_i) | 1 \leq i \leq n, 1 \leq j \leq m\}$ , the  $\mathcal{L}_{tri-d}$  for an instance  $x_i$  is calculated by:

$$\mathcal{L}_{tri-d} = \log \left( 1 + \exp(\|c_i - f_p^*\|_2^2 - \|c_i - f_n^*\|_2^2) \right), \quad (3)$$

where  $c_i$ ,  $f_p^*$ , and  $f_n^*$  indicate the anchor point, positive point, and negative point, respectively. Specifically,  $f_p^*$  and  $f_n^*$  is obtained by:

$$\begin{cases} f_p^* = \arg \max_{(f^*, y^*) \in \mathcal{F}_s, y^* = y_i} \|c_i - f^*\|_2^2 \\ f_n^* = \arg \min_{(f^*, y^*) \in \mathcal{F}_s, y^* \neq y_i} \|c_i - f^*\|_2^2 \end{cases}, \quad (4)$$

where  $(f^*, y^*)$  is the feature candidate and identity label pair.

Similarly, for each aligned image  $x'_i$ , we assign it the same identity as  $x_i$ . The instance variances of aligned data denoted as  $V'_t = \{(v'_i)^2\}_{i=1}^n$ , and centers, denoted as  $C'_t = \{c'_i\}_{i=1}^n$ , are then used to calculate the intra-distribution modeling losses according to (2), (3) and (4).

*Joint Modeling:* To guide the model in learning domain-unbiased knowledge, we integrate the new feature sampling set  $\mathcal{F}_s$  with the aligned feature sampling set  $\mathcal{F}'_s$ , creating the combined feature set denoted as  $\mathcal{F}_c = \mathcal{F}_s \cup \mathcal{F}'_s$ . Then, we utilize (3) and (4) to compute the joint modeling loss  $\mathcal{L}_{ce-j}$ , where  $c_i$  and  $c'_i$  serve as anchor points, and all sampled features are employed as positive and negative point candidates. Note that utilizing  $\mathcal{L}_{ce-j}$  and  $\mathcal{L}_{ce-d}$  simultaneously enables hierarchical modeling of both intra-domain and cross-domain distribution of the same identity, enhancing the knowledge mining capacity of the model.

3) *Prototype-based Knowledge Transfer*: To incorporate the knowledge encoded in the prototypes into the new model, we propose a Prototype Knowledge Transfer scheme: Firstly, as illustrated in Fig. 3(d), given a set of prototypes  $\mathcal{P}_{t-1} = \{(\mu_i, \sigma_i^2)\}_{i=1}^{N_{t-1}}$ , we sample  $N_{t-1}$  prototype features  $\mathcal{G}_{t-1} = \{g_i\}_{i=1}^{N_{t-1}}$  according to the distribution  $\mathcal{N}(\mu_i, \sigma_i^2)$ , where  $\mathcal{G}_{t-1}$  is formed as a matrix  $\mathbf{G}_{t-1} \in \mathbb{R}^{N_{t-1} \times d}$ . For a batch of  $n$  images where the predicted feature centers are transformed into a matrix  $\mathbf{F}_c \in \mathbb{R}^{n \times d}$ , we obtain a prototype-aware coordinate matrix  $\mathbf{C}_p$  by:

$$\mathbf{C}_p = \rho(\mathbf{F}_c \mathbf{G}_{t-1}^\top / \tau), \quad (5)$$

where each row of  $\mathbf{C}_p$  encodes the relative distance between  $x_i$  and all prototypes. Then, to ensure discrimination between the new and old data, while guiding the new model to effectively incorporate historical knowledge, we co-optimize the inter-instance affinity using the proposed Prototype Knowledge Transfer loss which is calculated by:

$$\mathcal{L}_{proto} = \mathcal{L}_{KL}(\rho(\mathbf{C}_p \mathbf{C}_p^\top / \tau) || \rho(\mathbf{F}_c \mathbf{F}_c^\top / \tau)). \quad (6)$$

where  $\rho(\mathbf{C}_p \mathbf{C}_p^\top)$  and  $\rho(\mathbf{F}_c \mathbf{F}_c^\top)$  are the inter-instance affinity matrix where each row represents the relative similarity between an instance  $x_i$  and all instances. Because  $\mathbf{C}_p$  is obtained by integrating old distribution knowledge with the new feature, optimizing (6) guarantees the acquisition of new data and mitigates forgetting of old knowledge.

#### D. Stage2: Post-Preparation

After completing the new LReID training, we train a new DTNet and new prototypes for future steps using the Distribution Transfer Training (DTT) and Distribution-oriented Prototype Generation (DPG) modules, respectively.

1) *Distribution Transfer Training*: As shown in Fig. 4(a), we begin by calculating the statistical information for each channel(R, G, B). Specifically, we compute the mean and standard deviation for the R, G, and B channels:  $(\mu_i^r, \sigma_i^r)$ ,  $(\mu_i^g, \sigma_i^g)$ , and  $(\mu_i^b, \sigma_i^b)$ . Then, the channel-wise statistics are utilized to augment the image distribution. Take the R channel as an example, the augmented mean  $\tilde{\mu}_i^r$  and standard deviation  $\tilde{\sigma}_i^r$  are sampled from the Gaussian distributions  $\mathcal{N}(\mu_i^r, (\sigma_i^r)^2)$  and  $\mathcal{N}(\sigma_i, (\sigma_i^r)^2)$ , respectively. Next, the augmented R channel data  $\tilde{x}_i^r$  is obtained via the intensity mapping function:

$$\tilde{x}_i^r = (x_i^r - \mu_i^r + \tilde{\mu}_i^r) \frac{\tilde{\sigma}_i^r}{\sigma_i^r}, \quad (7)$$

where  $x_i^r$  is the R channel of the input image  $x_i$ . Similarly, the augmented G channel  $\tilde{x}_i^g$  and B channel  $\tilde{x}_i^b$  can be generated following the above process. The resulting complete augmented image is denoted  $\tilde{x}_i$ .

Considering different instances show distinct distances from the old distribution, we propose to use a learnable network to transfer each individual instance into the target distribution. Specifically, the proposed DTNet  $\Phi_t$  consists of a lightweight encoder and a decoder. Given the input  $\tilde{x}_i$ , the generated image is  $\tilde{x}_i' = \Phi_t(\tilde{x}_i)$ . Then, a self-supervised reconstruction loss  $\mathcal{L}_{ReC}$

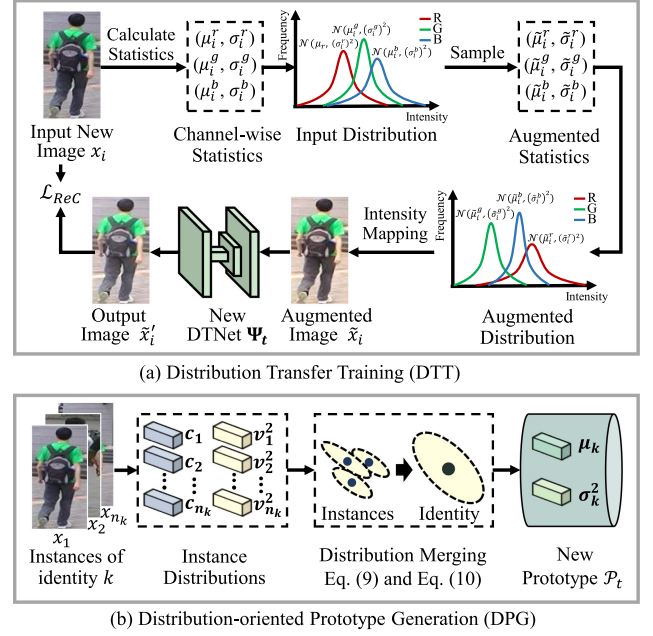


Fig. 4. (a) DTT aims to train a DTNet based on the new data utilizing the random distribution augmentation strategy. (b) DPG focuses on generating distribution-oriented prototypes for each identity.

is applied as follows:

$$\mathcal{L}_{ReC} = \|x_i - \tilde{x}_i'\|. \quad (8)$$

Since  $\tilde{x}_i$  is randomly generated, the model trained under the supervision of (8) can transfer the arbitrarily future data into the distribution of step  $t$ .

2) *Distribution-oriented Prototype Generation*: As shown in Fig. 4(b), to preserve abundant and informative knowledge of each identity without retaining any exemplar, the joint distribution of all instances within an identity  $k$  is formulated as the distribution-aware prototype, which is parameterized as  $\mathcal{N}(\mu_k, \sigma_k^2)$  where  $\mu_k \in \mathbb{R}^d$  and  $\sigma_k^2 \in \mathbb{R}^d$  represent the mean and variance respectively. Specifically, given identity  $k$  with  $n_k$  instances  $\{x_i^k\}_{i=1}^{n_k}$  in the training set (For simplicity, we denote these  $n_k$  instances as  $\{x_i\}_{i=1}^{n_k}$ ), the identity distribution center  $\mu_k$  could be obtained by:

$$\mu_k = \frac{1}{n_k} \sum_{i=1}^{n_k} c_i, \quad (9)$$

where  $c_i$  is the learned center of the instance  $x_i$ . Besides, the identity variance  $\sigma_k^2$  is calculated as follows:

$$\sigma_k^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} (c_i^2 + v_i^2) - \left( \frac{1}{n_k} \sum_{i=1}^{n_k} c_i \right)^2, \quad (10)$$

where

$$\varphi(x, \mu_k, \sigma_k) = \frac{e^{-\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)}}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \quad (11)$$

is the distribution function denoted by  $\mathcal{N}(\mu_k, \sigma_k^2)$  and  $\Sigma_k$  is a diagonal matrix whose diagonal elements are  $\sigma_k^2$ . Similarly,  $\varphi(x, c_i, v_i)$  is the distribution functions denoted by  $\mathcal{N}(c_i, v_i^2)$ .

More mathematical analysis of (9) and (10) is provided in Section IV. Compared to the statistical feature distribution-based prototypes [4] that only utilize  $c_i$  to obtain identity-level divergence, our proposed distribution-oriented prototype simultaneously considers  $c_i$  and  $v_i$  when calculating  $\sigma_k^2$ , making our prototype more informative.

#### E. Training and Inference

When training with the first dataset, the historical prototypes and DTNet are non-existent. Thus the LReID training loss is directly calculated by:

$$\mathcal{L} = \mathcal{L}_{ce-d} + \lambda_1 \mathcal{L}_{tri-d}, \quad (12)$$

where  $\lambda_1$  is the weight of  $\mathcal{L}_{tri-d}$ .

During the subsequent training steps, a basic version of our method is DKP, which is detailed in our conference paper [19]. The LReID loss function for DKP is defined as follows:

$$\mathcal{L} = \mathcal{L}_{ce-d} + \lambda_1 \mathcal{L}_{tri-d} + \lambda_2 \mathcal{L}_{proto}, \quad (13)$$

where only the new data are exploited and  $\lambda_2$  is a weight hyperparameter.

In our DKP++, we introduce a distribution aligning module into the DKP framework, achieving hierarchical modeling of both intra-domain and cross-domain distributions through the proposed loss functions. The overall loss for DKP++ is defined as follows:

$$\mathcal{L} = \mathcal{L}_{ce-d} + \lambda_1 \mathcal{L}_{tri-d} + \lambda_2 \mathcal{L}_{proto} + \lambda_3 \mathcal{L}_{ce-j} + \lambda_4 \mathcal{L}_{rel}, \quad (14)$$

where  $\lambda_3$  and  $\lambda_4$  are hyperparameters to balance the losses.

At the end of the  $t$ th learning step, to further blend the knowledge, we fuse the learned model  $M_t$  and the old model  $M_{t-1}$  as the final model

$$M_t \leftarrow (M_t + M_{t-1})/2. \quad (15)$$

During testing, we use the predicted feature center  $c$  of the input image  $x$  as shown in Fig. 3(a) for person matching.

### IV. THEORETICAL ANALYSIS AND JUSTIFICATIONS

1) *Justification of the Identity-level Distribution Generation Strategy:* For  $n_k$  instances  $\{x_i\}_{i=1}^{n_k}$  of identity  $k$ , the distribution of identity  $k$  is estimated by the mix of the distributions of  $\{x_i\}_{i=1}^{n_k}$  in this paper. So given the distribution function  $\varphi(x, c_i, v_i)$  denoted by  $\mathcal{N}(c_i, v_i^2)$ , the identity distribution function  $\varphi(x, \mu_k, \sigma_k)$  can be obtained by:

$$\varphi(x, \mu_k, \sigma_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \varphi(x, c_i, v_i), \quad (16)$$

where  $\mu_k$  and  $\sigma_k$  are the mean and standard deviation of the identity  $k$ . Besides, according to probability theory [61], we obtain  $c_i = \int x \varphi(x, c_i, v_i) dx$  and  $v_i^2 = \int x^2 \varphi(x, c_i, v_i) dx - (\int x \varphi(x, c_i, v_i) dx)^2$ . Moreover, the identity distribution center  $\mu_k$  could be calculated as follows:

$$\mu_k = \int x \varphi(x, \mu_k, \sigma_k) dx = \int \frac{1}{n_k} \sum_{i=1}^{n_k} x \varphi(x, c_i, v_i) dx$$

$$= \frac{1}{n_k} \sum_{i=1}^{n_k} \int x \varphi(x, c_i, v_i) dx = \frac{1}{n_k} \sum_{i=1}^{n_k} c_i \quad (17)$$

Similarly, the identity variance  $\sigma_k^2$  could be obtained by:

$$\begin{aligned} \sigma_k^2 &= \int x^2 \varphi(x, \mu_k, \sigma_k) dx - \left( \int x \varphi(x, \mu_k, \sigma_k) dx \right)^2 \\ &= \int x^2 \frac{1}{n_k} \sum_{i=1}^{n_k} \varphi(x, c_i, v_i) dx - \left( \frac{1}{n_k} \sum_{i=1}^{n_k} c_i \right)^2 \\ &= \frac{1}{n_k} \sum_{i=1}^{n_k} (c_i^2 + v_i^2) - \left( \frac{1}{n_k} \sum_{i=1}^{n_k} c_i \right)^2 \end{aligned} \quad (18)$$

Therefore, compared to the classical statistical variance  $\frac{1}{n_k} \sum_{i=1}^{n_k} c_i^2 - (\frac{1}{n_k} \sum_{i=1}^{n_k} c_i)^2$ , the generated identity variance extends it by integrating the learned instance-specific variances, which fill in the missing information of the input data.

2) *Theoretical Analysis of Our Distribution-oriented Prototype Compared to Existing Methods:* Given the actual identity distribution  $p(X) = \mathcal{N}(\mu, \Sigma)$ , which is a  $d$ -dimensional Gaussian distribution with the covariance matrix  $\Sigma = \text{diag}((\sigma^1)^2, (\sigma^2)^2, \dots, (\sigma^d)^2)$  and the mean  $\mu \in \mathbb{R}^d$ , let  $q(X) = \mathcal{N}(\hat{\mu}, \hat{\Sigma})$  represent the empirical Gaussian distribution based on  $n$  samples drawn from  $p(X)$ , where  $\hat{\mu}$  and  $\hat{\Sigma}$  are the estimated mean and covariance matrix, respectively. The Wasserstein-2 distance [62]  $W_2$  between  $p(X)$  and  $q(X)$  is given by:

$$\begin{aligned} W_2(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})) \\ = \text{Tr}(\hat{\Sigma} + \Sigma - 2(\hat{\Sigma}^{1/2} \Sigma \hat{\Sigma}^{1/2})^{1/2}) + \|\hat{\mu} - \mu\|^2. \end{aligned} \quad (19)$$

*Theorem 1:* When omitting the instance-level variances in (18), the identity variance generation process of our method degrades to statistical variance  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n f_i^2 - (\frac{1}{n} \sum_{i=1}^n f_i)^2$  and  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n f_i$ , where  $f_i$  is the feature extracted from  $X_i$ , i.e.,  $c_i$  in our paper. As  $n \rightarrow \infty$ , the Wasserstein-2 distance between  $p(X)$  and our prediction  $q_1(X)$  satisfies:

$$\mathbb{E}(W_2(p(X), q_1(X))) \rightarrow 0 \quad (20)$$

*Proof 1:* According to (19), the Wasserstein-2 distance between  $p(X)$  and our prediction  $q_1(X)$  is:

$$\begin{aligned} W_2(p(X), q_1(X)) &= \sum_{i=1}^d \left( (\mu^i - \hat{\mu}^i)^2 + (\hat{\sigma}^i)^2 \right. \\ &\quad \left. + (\sigma^i)^2 - 2\sigma^i \hat{\sigma}^i \right), \end{aligned} \quad (21)$$

where  $\sigma^i$  and  $\mu^i$  represent the  $i$ th element of the vectors  $\sigma$  and  $\mu$ , respectively. Since  $\hat{\mu}^i \sim \mathcal{N}(\mu^i, \frac{\sigma^i}{n})$ , we obtain  $\mathbb{E}((\mu^i - \hat{\mu}^i)^2) = \frac{(\sigma^i)^2}{n}$  and  $\mathbb{E}((\hat{\sigma}^i)^2) = \frac{(n-1)(\sigma^i)^2}{n}$ . According to the Cauchy Schwarz inequality [63], we obtain:

$$\mathbb{E}((\hat{\sigma}^i)^2) = \mathbb{E}(1\hat{\sigma}^i)^2 \leq \mathbb{E}(1^2)\mathbb{E}((\hat{\sigma}^i)^2) = \mathbb{E}((\hat{\sigma}^i)^2). \quad (22)$$

Therefore  $\mathbb{E}(\hat{\sigma}^i) \leq \sqrt{\frac{n-1}{n}}\sigma^i$  and then

$$\begin{aligned} \mathbb{E}(\mathbf{W}_2(p(X), q_1(X))) &= \sum_{i=1}^d (2(\sigma^i)^2 - 2\sigma^i\hat{\sigma}^i) \\ &\leq \sum_{i=1}^d 2 \left(1 - \sqrt{\frac{n-1}{n}}\right) \sigma^i = O\left(\frac{1}{n}\right) \end{aligned} \quad (23)$$

**Theorem 2:** When the actual distribution is anisotropic, as  $n \rightarrow \infty$ , the Wasserstein-2 distance between  $p(X)$  and PASS prediction  $q_2(X)$  satisfies:

$$\mathbb{E}(\mathbf{W}_2(p(X), q_1(X))) < \mathbb{E}(\mathbf{W}_2(p(X), q_2(X))) \quad (24)$$

**Proof2:** Existing methods [4], [42], e.g., PASS, introduced an isotropic variance where a scalar  $r = \frac{1}{d} \sum_{i=1}^d (\hat{\sigma}^i)^2$  is adopted as the variance prediction of each dimension. The Wasserstein-2 distance between  $p(X)$  and PASS et al. prediction  $q_2(X)$  is:

$$\mathbf{W}_2(p(X), q_2(X)) = \text{Tr}(rI + \Sigma - 2\sqrt{r\Sigma}) + \|\hat{\mu} - \mu\|^2 \quad (25)$$

Therefore, we obtain:

$$\mathbb{E}(\mathbf{W}_2(p(X), q_2(X))) = \sum_{i=1}^d \left(2(\sigma^i)^2 - 2\sigma^i\mathbb{E}(\sqrt{r}) + \frac{(\sigma^i)^2}{n}\right) \quad (26)$$

According to the Jensen's inequality [64], we obtain  $\mathbb{E}(\sqrt{r}) \leq \sqrt{\mathbb{E}(r)}$  since  $f(x) = \sqrt{x}$  is a concave function. Then  $\mathbb{E}(\sqrt{r}) \leq \sqrt{\frac{1}{d} \sum_{i=1}^d (\sigma^i)^2}$  (existing methods primarily use unbiased estimation of  $(\sigma^i)^2$  to calculate  $(\hat{\sigma}^i)^2$ ). Let  $A = \sum_{i=1}^d (\sigma^i)^2$ . According to the Cauchy Schwarz inequality,  $(\sum_{i=1}^d \sigma^i)^2 < dA$  (different  $\sigma^i$  are not always identical). Let  $\sqrt{dA} - \sum_{i=1}^d \sigma^i = \epsilon$ . Then we have:

$$\mathbb{E}(\mathbf{W}_2(p(X), q_2(X))) = 2\sqrt{\frac{A}{d}}\epsilon + \frac{(\sigma^i)^2}{n} \quad (27)$$

Thus, as  $n \rightarrow \infty$ ,  $\mathbb{E}(\mathbf{W}_2(p(X), q_2(X))) = 2\sqrt{\frac{A}{d}}\epsilon$ .

Therefore, the anisotropic variances introduced in this paper reduce the posterior risk compared to the isotropic variances employed in existing methods such as PASS [4] and FCS [42]. As illustrated in Fig. 5, the isotropic variances adopted by prototype-based methods such as PASS and FCS tend to introduce more outliers and ambiguous features. In contrast, the anisotropic variances used in our approach provide a more accurate characterization of the feature distribution.

## V. EXPERIMENTS

### A. Experimental Settings

**Benchmarks:** We conduct the experiments on the LReID benchmark [9], which comprises a total of twelve ReID datasets. Among them, five datasets (Market1501 [65], DukeMTMC-reID [66], CUHK-SYSU [67], MSMT17-V2 [21], and CUHK03 [68]) are used as seen domains for lifelong training

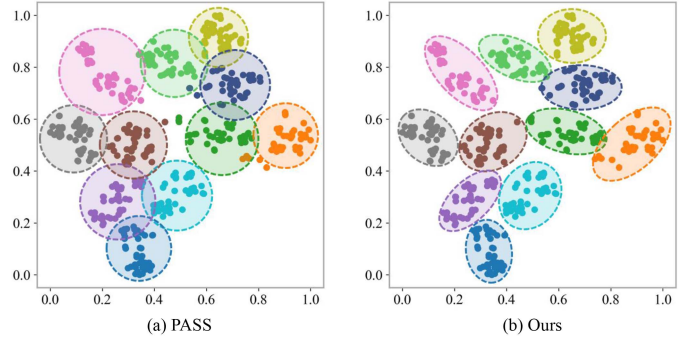


Fig. 5. Visualization of existing statistical feature distribution-based prototypes [4] and our instance-level distribution-aware prototypes.

and testing. Two different dataset orders are adopted to form various lifelong learning scenarios.<sup>1,2</sup> Additionally, the other seven datasets (CUHK01 [69], CUHK02 [70], VIPeR [71], PRID [72], i-LIDS [73], GRID [74], and SenseReID [75]) are tested as the unseen domains to show the generalization capacity of the models. The details of image and identity distributions for the above datasets are shown in Fig. 6, where “train” indicates the data used for lifelong training, while “query” and “gallery” are only used for evaluation.

**Evaluation Metrics:** The mean Average Precision (mAP) and Rank@1 accuracy (R@1) on each dataset are utilized to evaluate the model on specific domains. Additionally, the average mAP and average R@1 across all seen and unseen domains are calculated to compare different methods' overall lifelong learning capacity (on seen domains) and generalization capacity (on unseen domains), respectively.

**Implementation Details:** To ensure a fair comparison with prior LReID methods [9], [13], [14], we utilize the ResNet50 network [76] as our backbone. For all the training orders mentioned in this paper, the LReID model is trained on the first dataset for 80 epochs, and on subsequent datasets for 60 epochs each. A mini-batch size of 64 is adopted, where 16 identities are sampled with 4 images per identity during training. The model is optimized using the SGD optimizer with a learning rate of 0.008 and a weight decay of 0.0001. Furthermore, the loss weight parameters  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$ , and  $\lambda_4$  are set as 1.5, 1.5, 0.75, and 0.75, respectively. The temperature parameter  $\tau$  is set to 0.1. The instance sampling number  $m$  is set to 6 according to [19]. Instead of storing and exploiting all historical prototypes and DTNet models, only the  $t-1$  domain prototypes and DTNet model are utilized during  $t$  ( $t \geq 2$ ) domain training, thus the storage overhead of our method does not increase over time. The backbone of the DTNet is the MobileNet-v3 [77]. During DTNet training, the model optimization settings and image pre-processing are the same as those of the LReID model. All experiments are conducted on two NVIDIA 4090 GPUs under the PyTorch platform.

<sup>1</sup>(Order-1) Market-1501 → CUHK-SYSU → DukeMTMC → MSMT17 → CUHK03

<sup>2</sup>(Order-2) DukeMTMC → MSMT17 → Market-1501 → CUHK-SYSU → CUHK03

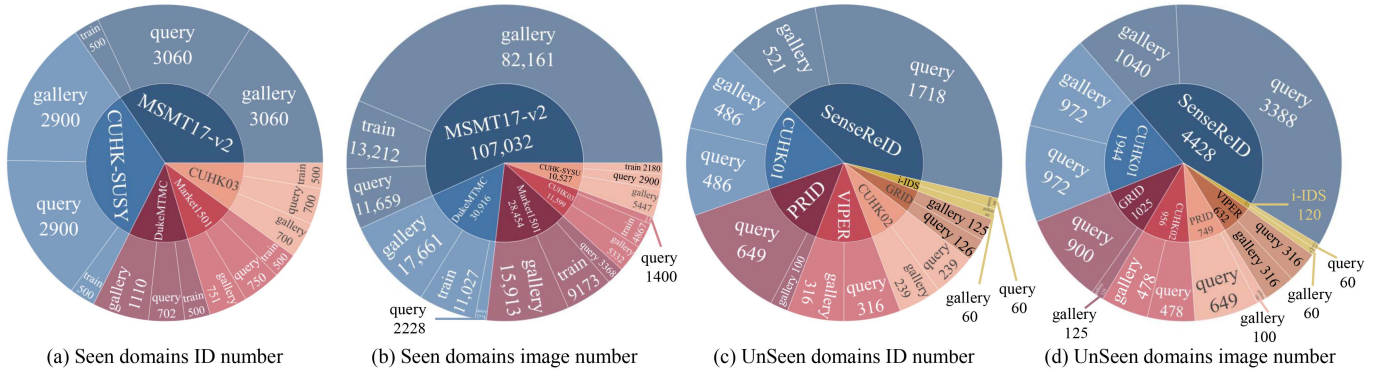


Fig. 6. The statistics of IDs and image numbers for seen and unseen domains in the LReID benchmark.

TABLE I  
PERFORMANCE COMPARISON ON TRAINING ORDER-1: MARKET-1501 → CUHK-SYSU → DUKEMTMC → MSMT17 → CUHK03

|       | Method                | Publication | Market-1501 |             | CUHK-SYSU   |             | DukeMTMC    |             | MSMT17      |             | 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         |
| CIL   | Joint-Train           | -           | 75.3        | 90.1        | 84.5        | 86.0        | 66.9        | 81.6        | 31.6        | 57.1        | 58.5        | 61.4        | 63.4        | 75.2        | 55.2        | 48.2        |
|       | LwF [78]              | T-PAMI 2017 | 56.3        | 77.1        | 72.9        | 75.1        | 29.6        | 46.5        | 6.0         | 16.6        | 36.1        | 37.5        | 40.2        | 50.6        | 47.2        | 42.6        |
|       | SPD [79]              | ICCV 2019   | 35.6        | 61.2        | 61.7        | 64.0        | 27.5        | 47.1        | 5.2         | 15.5        | 42.2        | 44.3        | 34.4        | 46.4        | 40.4        | 36.6        |
|       | PASS [4]              | CVPR 2021   | 60.7        | 79.4        | 81.1        | 83.0        | 41.0        | 57.0        | 12.7        | 28.7        | 29.9        | 27.4        | 45.1        | 55.1        | 54.5        | 47.2        |
|       | PRD [40]              | ICML 2023   | 7.3         | 18.0        | 33.5        | 35.6        | 3.7         | 7.6         | 0.8         | 2.4         | 33.8        | 33.8        | 15.8        | 19.5        | 23.0        | 17.7        |
|       | PRAKA [37]            | ICCV 2023   | 37.4        | 61.3        | 69.3        | 71.8        | 35.4        | 55.0        | 10.7        | 27.2        | <b>54.0</b> | <b>55.6</b> | 41.3        | 54.2        | 47.7        | 41.6        |
|       | FCS [42]              | CVPR 2024   | 58.3        | 78.5        | 75.1        | 76.2        | 42.6        | 59.8        | 10.2        | 24.3        | 35.3        | 34.9        | 44.3        | 54.7        | 52.1        | 44.2        |
| LReID | CRL [80]              | WACV 2021   | 58.0        | 78.2        | 72.5        | 75.1        | 28.3        | 45.2        | 6.0         | 15.8        | 37.4        | 39.8        | 40.5        | 50.8        | 47.8        | 43.5        |
|       | AKA [9]               | CVPR 2021   | 51.2        | 72.0        | 47.5        | 45.1        | 18.7        | 33.1        | 16.4        | 37.6        | 27.7        | 27.6        | 32.3        | 43.1        | 44.3        | 40.4        |
|       | AKA† [9]              | CVPR 2021   | 58.1        | 77.4        | 72.5        | 74.8        | 28.7        | 45.2        | 6.1         | 16.2        | 38.7        | 40.4        | 40.8        | 50.8        | 47.6        | 42.6        |
|       | PatchKD [13]          | ACM MM 2022 | <b>68.5</b> | <b>85.7</b> | 75.6        | 78.6        | 33.8        | 50.4        | 6.5         | 17.0        | 34.1        | 36.8        | 43.7        | 53.7        | 49.1        | 45.4        |
|       | ConRFL [35]           | PR 2023     | 59.2        | 78.3        | 82.1        | 84.3        | 45.6        | 61.8        | 12.6        | 30.4        | <b>51.7</b> | <b>53.8</b> | 50.2        | 61.7        | 57.4        | <b>52.3</b> |
|       | MEGE [81]             | T-PAMI 2023 | 39.0        | 61.6        | 73.3        | 76.6        | 16.9        | 30.3        | 4.6         | 13.4        | 36.4        | 37.1        | 34.0        | 43.8        | 47.7        | 44.0        |
|       | CKP [15]              | ACM MM 2024 | 53.8        | 76.0        | 81.2        | 83.0        | 49.7        | 67.0        | 18.4        | 40.8        | 44.1        | 45.8        | 49.4        | 62.5        | 58.0        | 51.0        |
|       | LSTKC [14]            | AAAI 2024   | 54.7        | 76.0        | 81.1        | 83.4        | 49.4        | 66.2        | <b>20.0</b> | <b>43.2</b> | 44.7        | 46.5        | 50.0        | 63.1        | 57.0        | 49.9        |
|       | C <sup>2</sup> R [31] | CVPR 2024   | <b>69.0</b> | <b>86.8</b> | 76.7        | 79.5        | 33.2        | 48.6        | 6.6         | 17.4        | 35.6        | 36.2        | 44.2        | 53.7        | 48.7        | 41.9        |
|       | DKP                   | This Paper  | 60.3        | 80.6        | <b>83.6</b> | <b>85.4</b> | <b>51.6</b> | <b>68.4</b> | 19.7        | 41.8        | 43.6        | 44.2        | <b>51.8</b> | <b>64.1</b> | <b>59.2</b> | 51.6        |
|       | DKP++                 | This Paper  | 63.9        | 83.6        | <b>83.7</b> | <b>85.3</b> | <b>56.7</b> | <b>74.1</b> | <b>26.5</b> | <b>52.6</b> | 46.5        | 47.6        | <b>55.4</b> | <b>68.6</b> | <b>65.7</b> | <b>58.5</b> |

† indicates the results are re-implemented by [13].

## B. The Compared Methods

To extensively evaluate our method, various state-of-the-art non-exemplar LReID approaches including CRL [80], AKA [9], PatchKD [13], ConRFL [35], MEGE [81], CKP [15], LSTKC [14] and C<sup>2</sup>R [31] are compared. Additionally, the latest class incremental learning (CIL) methods, LwF [78], SPD [82], PASS [4], PRD [40], PRAKA [37], and FCS [42]), are also tested. To ensure a fair comparison, all models are implemented with the same backbone and training settings. Thus for the CIL models, we incorporate the widely adopted triplet loss [23] to make a fair comparison with the LReID methods. In addition, the *Joint-Train* result, the upper bound of the LReID models where all datasets are given at once for training, is also reported.

## C. Seen-Domain Performance Evaluation

We present the results of different methods on each seen domain and the average performance across all seen domains (Seen-Avg) in Tables I and II, corresponding to Training Order-1 and Training Order-2, respectively. The best results are highlighted in **RED**, while the second-best results are marked in **BLUE**.

*Compared to LReID Methods:* As reported in Tables I and II, our base version DKP demonstrates competitive performance against various state-of-the-art LReID methods across individual datasets. Specifically, DKP achieves an improvement of **1.6%/1.0%** and **1.6%/1.1%** in Seen-Avg mAP/R@1 under two different training orders compared to the state-of-the-art methods. These results arise because existing methods either suffer from inadequate anti-forgetting capacity or limited new knowledge acquisition ability. For example, C<sup>2</sup>R and PatchKD achieve leading performance on the first dataset, benefiting from robust person-relevant knowledge distillation constraints. However, these constraints also significantly hinder the model from learning new knowledge. Besides, ConRFL performs well on the last dataset, but its performance on the earlier datasets is significantly limited, especially on Training Order-2, as shown in Table II. In contrast, our method achieves strong new knowledge acquisition and anti-forgetting simultaneously, thanks to the proposed distribution prototyping mechanism, which precisely retains historical information to facilitate new knowledge mining and old knowledge consolidation.

Moreover, our DKP++ further achieves at least **5.2%/5.5%** and **5.8%/7.0%** Seen-Avg mAP/R@1 improvements under two

TABLE II  
PERFORMANCE COMPARISON ON TRAINING ORDER-2: DUKEMTMC → MSMT17 → MARKET-1501 → CUHK-SYSU → CUHK03

|       | Method                | Publication | DukeMTMC    |             | MSMT17      |             | Market-1501 |             | 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         |
| CIL   | Joint-Train           | -           | 66.9        | 81.6        | 31.6        | 57.1        | 75.3        | 90.1        | 84.5        | 86.0        | 58.5        | 61.4        | 63.4        | 75.2        | 55.2        | 48.2        |
|       | LwF [78]              | T-PAMI 2017 | 42.7        | 61.7        | 5.1         | 14.3        | 34.4        | 58.6        | 69.9        | 73.0        | 34.1        | 34.1        | 37.2        | 48.4        | 44.0        | 40.1        |
|       | SPD [79]              | ICCV 2019   | 28.5        | 48.5        | 3.7         | 11.5        | 32.3        | 57.4        | 62.1        | 65.0        | 43.0        | 45.2        | 33.9        | 45.5        | 39.8        | 36.3        |
|       | PASS [4]              | CVPR 2021   | 57.6        | 72.6        | 12.9        | 29.9        | 49.2        | 71.9        | 81.5        | 83.2        | 30.5        | 29.1        | 46.3        | 57.3        | 54.5        | 47.3        |
|       | PRD [40]              | ICML 2023   | 3.6         | 8.2         | 0.6         | 1.8         | 8.9         | 22.3        | 34.6        | 36.1        | 35.4        | 35.3        | 16.6        | 20.7        | 20.7        | 15.0        |
|       | PRAKA [37]            | ICCV 2023   | 31.2        | 48.7        | 6.6         | 19.1        | 47.8        | 69.8        | 70.4        | 73.0        | <b>54.9</b> | <b>56.6</b> | 42.2        | 53.4        | 48.4        | 41.1        |
|       | FCS [42]              | CVPR 2024   | 53.6        | 70.0        | 9.5         | 23.5        | 48.7        | 69.9        | 76.2        | 78.2        | 37.1        | 38.4        | 45.0        | 56.0        | 52.7        | 45.1        |
| LReID | CRL [80]              | WACV 2021   | 43.5        | 63.1        | 4.8         | 13.7        | 35.0        | 59.8        | 70.0        | 72.8        | 34.5        | 36.8        | 37.6        | 49.2        | 45.3        | 41.4        |
|       | AKA [9]               | CVPR 2021   | 32.5        | 49.7        | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | 40.8        | 37.2        |
|       | AKA† [9]              | CVPR 2021   | 42.2        | 60.1        | 5.4         | 15.1        | 37.2        | 59.8        | 71.2        | 73.9        | 36.9        | 37.9        | 38.6        | 49.4        | 46.0        | 41.7        |
|       | PatchKD [13]          | ACM MM 2022 | <b>58.3</b> | <b>74.1</b> | 6.4         | 17.4        | 43.2        | 67.4        | 74.5        | 76.9        | 33.7        | 34.8        | 43.2        | 54.1        | 48.6        | 44.1        |
|       | ConRFL [35]           | PR 2023     | 34.4        | 51.3        | 7.6         | 20.1        | <b>61.6</b> | 80.4        | 82.8        | <b>85.1</b> | 49.0        | 50.1        | 47.1        | 57.4        | 57.9        | <b>53.4</b> |
|       | MEGE [81]             | T-PAMI 2023 | 21.6        | 35.5        | 3.0         | 9.3         | 25.0        | 49.8        | 69.9        | 73.1        | 34.7        | 35.1        | 30.8        | 40.6        | 44.3        | 41.1        |
|       | CKP [15]              | ACM MM 2024 | 49.4        | 67.0        | 14.5        | 33.8        | 56.0        | 77.6        | <b>83.2</b> | 84.9        | 45.3        | 47.1        | 49.7        | 62.1        | 57.2        | 50.0        |
|       | LSTKC [14]            | AAAI 2024   | 49.9        | 67.6        | <b>14.6</b> | <b>34.0</b> | 55.1        | 76.7        | 82.3        | 83.8        | 46.3        | 48.1        | 49.6        | 62.1        | 57.6        | 49.6        |
|       | C <sup>2</sup> R [31] | CVPR 2024   | <b>59.7</b> | <b>75.0</b> | 7.3         | 19.2        | 42.4        | 66.5        | 76.0        | 77.8        | 37.8        | 39.3        | 44.7        | 55.6        | 48.5        | 41.7        |
|       | <b>DKP</b>            | This Paper  | 53.4        | 70.5        | 14.5        | 33.3        | 60.6        | <b>81.0</b> | 83.0        | 84.9        | 45.0        | 46.1        | <b>51.3</b> | <b>63.2</b> | <b>59.0</b> | 51.6        |
|       | <b>DKP++</b>          | This Paper  | 55.7        | 73.9        | <b>22.2</b> | <b>47.8</b> | <b>65.9</b> | <b>85.6</b> | <b>84.4</b> | <b>86.5</b> | <b>49.5</b> | <b>51.7</b> | <b>55.5</b> | <b>69.1</b> | <b>64.9</b> | <b>57.9</b> |

† indicates the results are re-implemented by [13]. — represents the results that are not reported in the official papers.

training orders compared to the existing methods. In particular, the performance on the latest datasets is improved significantly compared to DKP. These results are attributed to our distribution-aligning design improves the utilization of prototypes, resulting in enhanced new knowledge learning capacity.

*Compared to CIL Methods:* Our DKP outperforms the CIL methods on the CUHK-SYSU, DukeMTMC, and MSMT17 datasets, exhibiting an average mAP/R@1 improvement of 6.7%/9.0% and 5.0%/5.9% across both training orders for seen domains. Our DKP++ model further achieves 10.3%/13.5% and 9.2%/11.8% Seen-Avg mAP/R@1 improvement. We observe that although the prototype-based model PRAKA excels our DKP/DKP++ on the last domain CUHK03, our model achieves remarkably higher performance than PRAKA in the early-stage domains and superior average performance. owing to the abundant information encoded by our distribution-aware prototypes, which mitigates the forgetting caused by domain shifts. Besides, the other prototype-based methods PASS and FCS obtain comparable performance with ours on the first dataset but suffer from inadequate performance on subsequent datasets. In contrast, our proposed prototype-based knowledge transfer scheme introduces adaptive feature knowledge distillation, which not only enhances the retention of old knowledge but also minimizes interference in the learning of new knowledge.

*Seen Domain Performance Tendency:* To verify the anti-forgetting capacity of different models, we evaluate the LReID model after each training step. The results are illustrated in Fig. 7. Initially, our model shows slightly inferior performance compared to existing methods. This initial disparity can be attributed to our proposed instance-level distribution learning component, which partially slows down convergence. However, as training progresses, our model demonstrates superior performance. This improvement is primarily due to our distribution prototyping design, which enhances the utilization of historical information during the learning of new knowledge, thereby improving long-term consolidation capacity. Additionally, the distribution aligning scheme in our DKP++ facilitates better

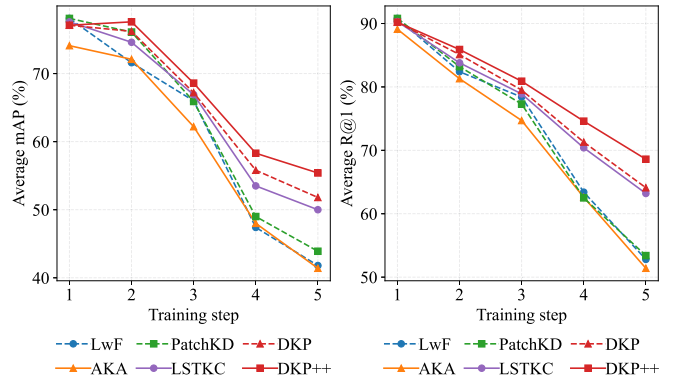


Fig. 7. Performance tendency on seen domains. After each training step, the model is evaluated on the already-seen domains.

interaction between new and old knowledge, contributing to further enhancements in model performance over time.

#### D. Unseen-Domain Generalization Evaluation

The average performance on unseen domains is shown in the last two columns of Tables I and II. Our DKP demonstrates superior performance compared to state-of-the-art CIL models, exhibiting an Unseen-Avg mAP/R@1 improvement of 4.7%/4.4% and 4.5%/4.3% across both training orders. Furthermore, DKP also achieves competitive results with the state-of-the-art LReID method ConRFL. Note that ConRFL stores all historical LReID models to conduct knowledge distillation, thus showing superior performance compared to the other competitors. In contrast, our DKP method only stores a few historical prototypes, incurring negligible overhead. Moreover, DKP++ employs a lightweight DTNet model to facilitate knowledge transfer, outperforming ConRFL by **8.3%/6.2%** and **7.0%/4.5%** Unseen-Avg mAP/R@1 under two training orders, with less storage overhead. These results show that our knowledge aligning and

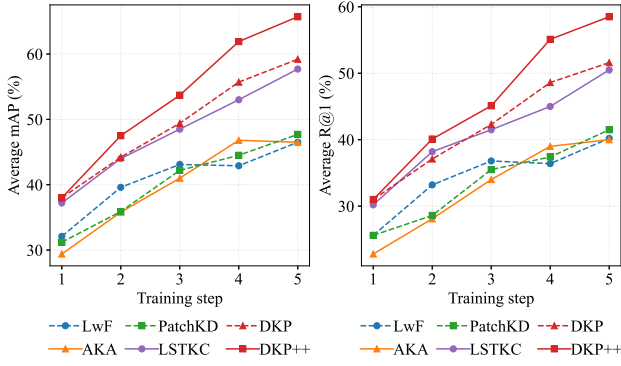


Fig. 8. Performance tendency on unseen domains. After each training step, the performance of all unseen domains is evaluated.

TABLE III  
ABLATION STUDY OF DIFFERENT LReID MODEL COMPONENTS

| Method   | DAM | IDM | PKT | Seen-Avg<br>mAP | Seen-Avg<br>R@1 | Unseen-Avg<br>mAP | Unseen-Avg<br>R@1 |
|----------|-----|-----|-----|-----------------|-----------------|-------------------|-------------------|
| Baseline |     |     |     | 44.0            | 53.4            | 46.8              | 39.8              |
|          | ✓   |     |     | 48.4            | 61.4            | 55.5              | 48.6              |
|          |     | ✓   |     | 46.1            | 60.0            | 51.6              | 45.5              |
|          |     |     | ✓   | 50.0            | 62.2            | 57.7              | 49.9              |
| DKP      |     | ✓   | ✓   | <b>51.8</b>     | <b>64.1</b>     | <b>59.2</b>       | <b>51.6</b>       |
| DKP++    | ✓   | ✓   | ✓   | <b>55.4</b>     | <b>68.6</b>     | <b>65.7</b>       | <b>58.5</b>       |

prototyping designs can effectively store and consolidate generalizable knowledge.

**Generalization Curves:** We further analyze the average performance on unseen domains throughout the lifelong training steps, as illustrated in Fig. 8. Both DKP and DKP++ demonstrate comparable performance to the state-of-the-art LSTKC at the initial training step. However, as the training steps increase, DKP consistently outperforms LSTKC due to the effective inter-domain information retention capacity of our distribution-oriented prototype modeling and utilization mechanism. Moreover, DKP++ shows a marked improvement over both LSTKC and DKP as training steps increase. This enhanced performance can be attributed to the distribution alignment mechanism in DKP++, which encourages the model to focus on cross-domain generalizable knowledge that improves the model's adaptability to diverse unknown domains. In contrast, existing methods tend to focus on domain-relevant information, resulting in the storage of substantial knowledge that is overfit to seen domains.

### E. Ablation Studies

**Influence of Different Model Components:** As illustrated in Fig. 3, the learning process of the LReID model is directly influenced by three key modules: DAM, IDM, and PKT. DTT and DPG serve as auxiliary components that prepare the models and prototypes for DAM and PKT, respectively. Thus, we focus our ablation studies on DAM, IDM, and PKT. Specifically, as shown in Table III, the *Baseline* refers to the DKP++ framework without the DAM, IDM, and PKT modules. We incrementally introduce these modules to assess their individual and combined contributions to performance. As observed, the inclusion of each module leads to consistent improvements across both seen and

unseen domains. The cross-domain distribution aligning capability of DAM effectively mitigates catastrophic forgetting by reducing domain gaps. IDM's instance-level distribution modeling enhances the model's ability to learn more discriminative features. Additionally, PKT promotes the transfer of distribution knowledge encoded in prototypes, which significantly reduces forgetting. When IDM and PKT are used together in the DKP model, performance is further enhanced, as the instance distributions modeled by IDM boost the knowledge transfer effectiveness of PKT. Finally, we achieve further performance gains by integrating DAM, IDM, and PKT into the complete DKP++ model. This integration enhances cross-domain connections through distribution alignment, which strengthens prototype-based knowledge transfer. Additionally, the IDM module is optimized to learn more robust distributions, as the domain-level information within each new instance is enriched.

**Influence of Hyperparameters:** The primary hyperparameters in this paper are the weights of the different loss components, whose effects are visualized in Fig. 9. Specifically, Fig. 9(a)~(d) illustrate the model performance as the hyperparameters  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$ , and  $\lambda_4$  vary. In Fig. 9(a), it is evident that model performance is relatively sensitive to changes in  $\lambda_1$ . This sensitivity arises because  $\mathcal{L}_{tri-d}$  effectively guides the model in learning discriminative features. However, an excessively large value for  $\lambda_1$  can cause optimization instability. In contrast, in Fig. 9(b),  $\lambda_2$  shows a more stable influence on model performance since  $\mathcal{L}_{proto}$  introduces a soft constraint to the model where the prototype and new features are adaptively distinguished. In Fig. 9(c), we observe that  $\lambda_3$  primarily influences the unseen domain performance. This is because  $\mathcal{L}_{tri-j}$  forces the model to mine the cross-domain generalizable knowledge which enhances the model's adaptivity to unknown domains. In Fig. 9(d),  $\lambda_4$  also shows a significant impact on the model performance since  $\mathcal{L}_{rel}$  is effective in maintaining historical knowledge by constricting the output distribution of the aligned data. Based on the results in Fig. 9, the default values are set to  $\lambda_1 = 1.5$ ,  $\lambda_2 = 1.5$ ,  $\lambda_3 = 0.75$ ,  $\lambda_4 = 0.75$  for the experiments in this paper.

**Investigation on Preserved DTNet Model Number:** In this paper, we retain only the DTNet model and prototypes from the  $(t-1)$ th domain in our DKP++ framework to prevent linearly increasing storage overhead. Additionally, we investigate a variant, denoted as DKP++<sup>†</sup>, which preserves all DTNet models and prototypes from historical domains. As shown in Table V, DKP++<sup>†</sup> achieves **0.2%/0.3%** improvements in Seen-Avg mAP/R@1 over DKP++, demonstrating that maintaining multiple DTNets and prototypes enhances knowledge consolidation. However, DKP++<sup>†</sup> suffers a **1.2%/0.6%** drop in UnSeen-Avg mAP/R@1 compared to DKP++. This degradation is primarily due to representation inconsistency caused by frequent DTNet switching during training, which impairs generalization capacity. Therefore, the single-DTNet version (DKP++) provides a better trade-off between storage cost, knowledge retention, and generalization capability.

**Investigation on Diverse Training Orders:** To comprehensively assess the effectiveness of the proposed method across diverse domain gaps, we conduct experiments on five additional dataset orders following the protocol of MEGE [81]. The

TABLE IV  
PERFORMANCE COMPARISON UNDER DIFFERENT PRE-TRAINED MODELS

| Pretrain | Method       | Publication | Market-1501 |             | CUHK-SYSU   |             | DukeMTMC    |             | MSMT17      |             | 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         |
| VF-Res   | LSTKC        | AAAI 2024   | <b>74.0</b> | <b>89.9</b> | 84.3        | 85.7        | 60.7        | 74.4        | 20.5        | 41.7        | 53.5        | 54.6        | 58.6        | 69.3        | 67.2        | 60.5        |
|          | CKP          | ACM MM 2024 | 61.7        | 82.0        | <b>89.9</b> | <b>91.3</b> | 60.7        | 75.1        | 26.5        | 48.4        | 53.1        | 53.4        | 58.4        | 70.0        | <b>70.8</b> | <b>63.4</b> |
|          | <b>DKP</b>   | This Paper  | 73.4        | 87.6        | <b>86.2</b> | <b>87.7</b> | <b>62.1</b> | <b>76.1</b> | <b>26.6</b> | <b>49.8</b> | <b>61.3</b> | <b>63.7</b> | <b>61.9</b> | <b>73.0</b> | 68.4        | 61.5        |
|          | <b>DKP++</b> | This Paper  | <b>74.7</b> | <b>89.4</b> | 83.7        | 84.3        | <b>63.6</b> | <b>78.8</b> | <b>33.2</b> | <b>60.0</b> | <b>59.5</b> | <b>60.6</b> | <b>62.9</b> | <b>74.6</b> | <b>71.4</b> | <b>64.1</b> |
| VF-ViT   | LSTKC        | AAAI 2024   | 63.4        | 81.6        | 82.2        | 84.3        | 57.9        | 73.5        | 21.5        | 43.3        | 51.0        | 53.0        | 55.2        | 67.1        | 62.7        | 55.3        |
|          | CKP          | ACM MM 2024 | 69.9        | <b>86.6</b> | <b>88.0</b> | <b>89.3</b> | 61.0        | 75.0        | <b>30.0</b> | <b>55.3</b> | 52.7        | 52.9        | 60.3        | 71.8        | 66.3        | 58.4        |
|          | <b>DKP</b>   | This Paper  | <b>70.6</b> | 86.1        | 84.7        | 86.2        | <b>61.8</b> | <b>77.2</b> | 29.2        | 53.5        | <b>58.9</b> | <b>61.4</b> | <b>61.0</b> | <b>72.9</b> | <b>67.3</b> | <b>59.6</b> |
|          | <b>DKP++</b> | This Paper  | <b>75.8</b> | <b>89.0</b> | <b>90.1</b> | <b>91.4</b> | <b>66.3</b> | <b>81.0</b> | <b>34.7</b> | <b>61.0</b> | <b>60.6</b> | <b>61.6</b> | <b>65.5</b> | <b>76.8</b> | <b>73.1</b> | <b>66.0</b> |

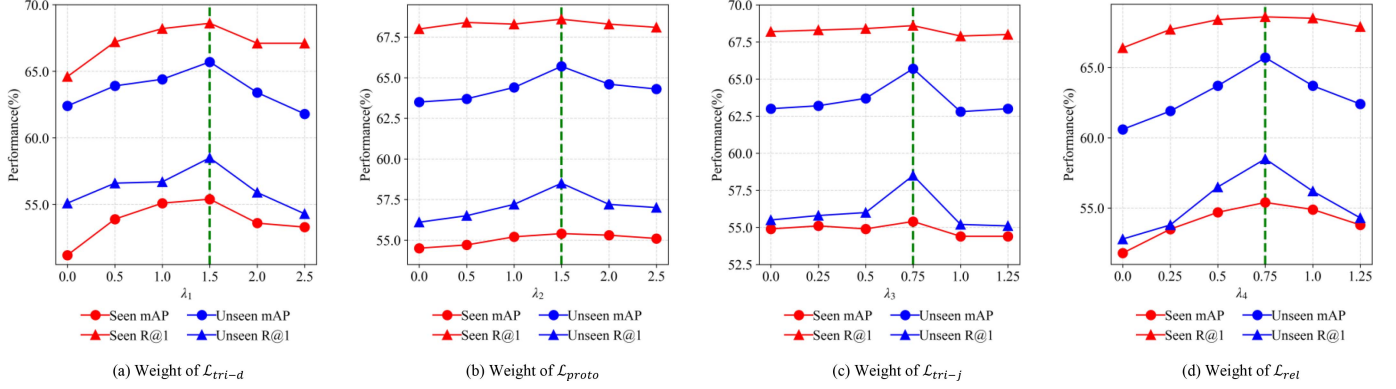


Fig. 9. Ablation studies on hyperparameters. The values marked by the dashed lines are adopted by our proposed method.

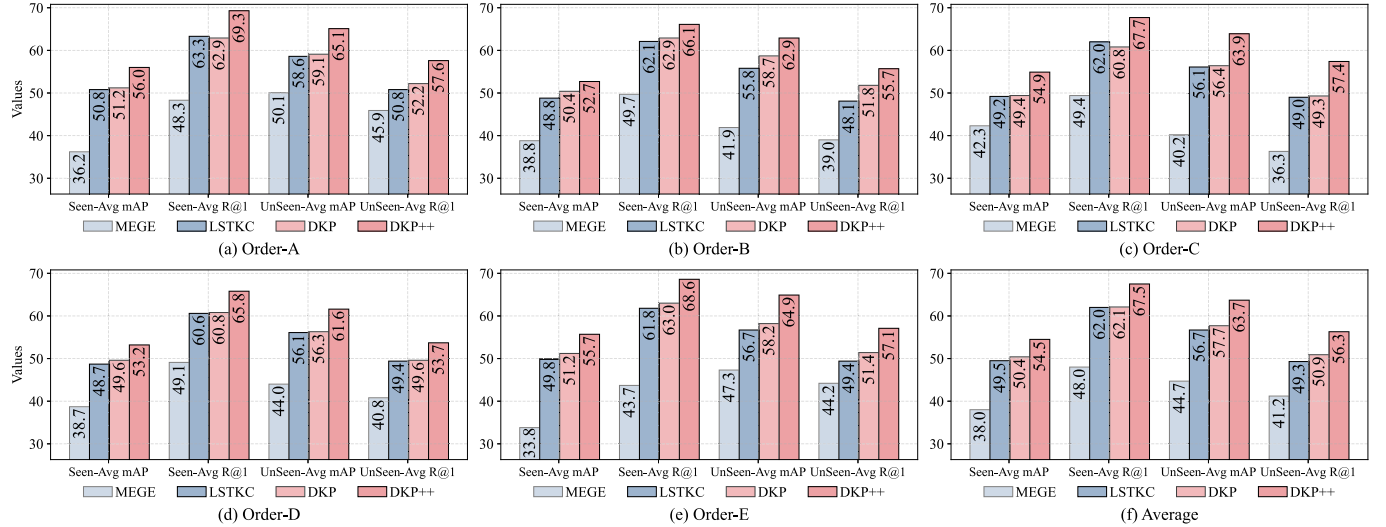


Fig. 10. Performance comparison on 5 additional training orders. (a) Order-A: MS  $\rightarrow$  SY  $\rightarrow$  DU  $\rightarrow$  MA  $\rightarrow$  CU. (b) Order-B: DU  $\rightarrow$  MA  $\rightarrow$  CU  $\rightarrow$  MS  $\rightarrow$  SY. (c) Order-C: SY  $\rightarrow$  DU  $\rightarrow$  CU  $\rightarrow$  MS  $\rightarrow$  MA. (d) Order-D: CU  $\rightarrow$  MS  $\rightarrow$  DU  $\rightarrow$  MA  $\rightarrow$  SY. (e) Order-E: MA  $\rightarrow$  MS  $\rightarrow$  DU  $\rightarrow$  SY  $\rightarrow$  CU. (f) The average performance of the above five orders. MS, SY, DU, MA, and CU are abbreviations for MSMT17, CUHK-SYSU, DukeMTMC, Market-1501, and CUHK03, respectively.

TABLE V  
COMPARISON OF PRESERVING SINGLE AND MULTIPLE DTNET MODELS

| Method                   | Type           | Seen-Avg    |             | Unseen-Avg  |             |
|--------------------------|----------------|-------------|-------------|-------------|-------------|
|                          |                | mAP         | R@1         | mAP         | R@1         |
| <b>DKP++</b>             | Single DTNet   | <b>55.4</b> | <b>68.6</b> | <b>65.7</b> | <b>58.5</b> |
| <b>DKP++<sup>†</sup></b> | Multiple DTNet | <b>55.6</b> | <b>68.9</b> | <b>64.5</b> | <b>57.1</b> |

results are presented in Fig. 10(a)~(e). Our DKP outperforms the state-of-the-art LSTKC method on three out of five orders. Moreover, DKP++ achieves superior performance across all

training orders. In Fig. 10(f), we present the average performance across the five training orders. DKP demonstrates **0.9%/0.1%** and **1.0%/1.6%** average mAP/R@1 improvement over LSTKC on the seen and unseen domains, respectively. Additionally, DKP++ surpasses LSTKC by **5.0%/5.5%** and **7.0%/7.0%** on Seen-Avg mAP/R@1 and UnSeen-Avg mAP/R@1 metrics, respectively. These results highlight the robustness of the proposed DKP to varying domain gaps, attributable to the domain aligning mechanism, which enhances cross-domain knowledge compatibility.

TABLE VI  
COMPARISON OF DIFFERENT FEATURE AND IMAGE-LEVEL STYLE TRANSFER METHODS

| Align | Method                     | Seen-Avg    |             | Unseen-Avg  |             |
|-------|----------------------------|-------------|-------------|-------------|-------------|
|       |                            | mAP         | R@1         | mAP         | R@1         |
| Feat  | PTKP                       | 50.7        | 63.6        | 58.8        | 51.9        |
|       | <b>DKP</b>                 | <b>51.8</b> | <b>64.1</b> | 59.2        | 51.6        |
| Img   | DKP + UnityGAN             | 48.1        | 62.9        | <b>60.2</b> | <b>52.8</b> |
|       | <b>DKP + DTNet (DKP++)</b> | <b>55.4</b> | <b>68.6</b> | <b>65.7</b> | <b>58.5</b> |

### F. Further Discussion With Existing Methods

*Adaptation to Vision Foundation Models:* Recently, a series of vision foundation models tailored for person re-identification have been proposed to enhance ReID performance [83], [84], [85], [86]. To evaluate the adaptability of our method to these stronger backbones, we compare with state-of-the-art LReID approaches, LSTKC and CKP, using the latest open-source ResNet-based (VF-Res) [84] and ViT-based (VF-ViT) [85] pre-trained models. As shown in Table IV, all methods benefit from these enhanced backbones when compared to the ImageNet-pretrained models reported in Table I.

With VF-Res as the backbone, DKP achieves at least **3.3%/3.0%** gains in Seen-Avg mAP/R@1 over LSTKC and CKP. DKP++ further improves performance, surpassing both baselines by **4.3%/4.6%** in Seen-Avg mAP/R@1, demonstrating superior knowledge consolidation from previously seen domains. While CKP exhibits slightly higher or comparable performance to DKP and DKP++ in terms of Unseen-Avg metrics, benefiting from its noise-robust design that mitigates the impact of hard samples, our DKP++ still achieves **0.6%/0.7%** improvements in Unseen-Avg mAP/R@1, owing to its enhanced domain adaptability via the proposed distribution alignment module.

When using VF-ViT as the backbone, DKP maintains consistent advantages, with at least **0.7%/1.1%** and **1.0%/1.2%** improvements over CKP and LSTKC in Seen-Avg and Unseen-Avg mAP/R@1, respectively. Remarkably, DKP++ delivers substantial performance gains across all domains, achieving at least **5.2%/5.0%** in Seen-Avg and **6.8%/7.6%** in Unseen-Avg mAP/R@1. These results validate the effectiveness of our distribution modeling and alignment mechanisms in lifelong learning scenarios.

In summary, vision foundation models offer a strong baseline for LReID. Built upon them, our method consistently delivers further improvements by integrating style alignment and distribution-aware knowledge consolidation strategies.

*Comparison of different style transfer methods:* We have conducted experiments to compare with the feature space alignment method PTKP [8] and the image style transfer method UnityGAN [51]. When re-implementing PTKP, the non-exemplar setting is adopted for a fair comparison. As shown in Table VI, when compared to PTKP, our DKP exhibits superior performance on Seen-Avg mAP/R@1. These improvements are achieved because our prototypes achieve a more accurate old data distribution description compared to the reserved feature distribution in PTKP.

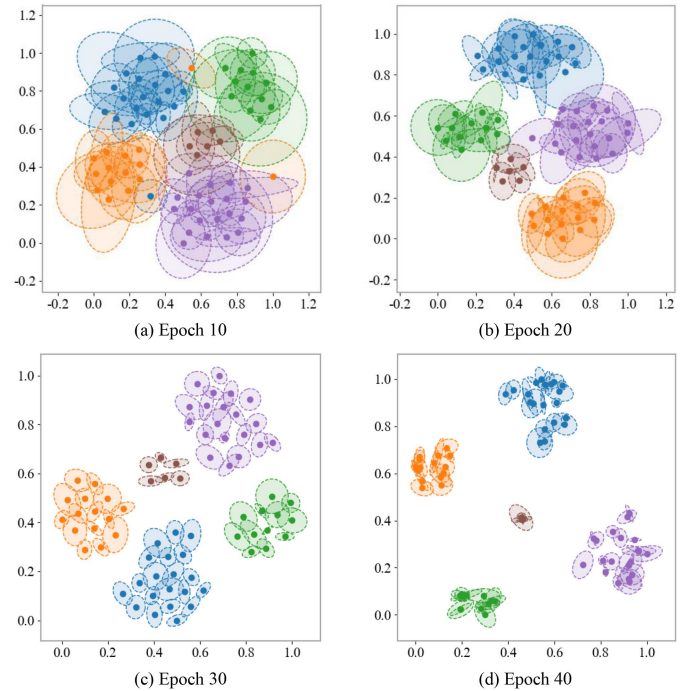


Fig. 11. The visualization of the learned instance-level distributions under different training epochs.

When comparing with UnityGAN, we replace our DTNet module with UnityGAN in the DKP++ framework. This substitution results in degraded performance in terms of Seen-Avg mAP/R@1 compared to the original DKP baseline. These results are primarily due to the limited training data available in each domain, which is insufficient to train a robust deep generative model. As a result, UnityGAN struggles to generate high-quality images representative of the historical domains. In contrast, our DTNet-based DKP++ model achieves substantial improvements of **7.3%/5.7%** in Seen-Avg mAP/R@1 and **5.5%/5.7%** in Unseen-Avg mAP/R@1 over the UnityGAN-based variant. These results highlight the superior robustness of our approach, which benefits from higher-quality domain-transferred data, thereby enhancing the preservation and utilization of historical knowledge.

### G. Visualization Results

1) *Instance-specific Distribution Learning Process:* To illustrate the distribution modeling process of our method, t-SNE visualizations are presented in Fig. 11, showcasing the evolution of learned centers and variances of training samples across different epochs. Initially, the features are widely dispersed, exhibiting large variances. As training progresses, the features become more tightly clustered, and the corresponding variances decrease. Simultaneously, the discriminative boundaries between different clusters are obtained, indicating that different individuals are better distinguished.

2) *Effect of Distribution Aligning:* To illustrate the effectiveness of our distribution alignment strategy on input images, we make a visual comparison in Fig. 12 against the existing image style transfer method UnityGAN [51]. UnityGAN



Fig. 12. Influence of distribution alignment on images. Compared to the existing method UnityGAN, the input images are transformed into the target domain styles by our distribution aligning module, with the person-aware information effectively preserved.

tends to introduce noticeable distortions in image details due to its reliance on a deep generative network, which requires large-scale data to achieve robustness. In the LReID setting, where training data is inherently limited, such reliance results in suboptimal generalization. In contrast, our image style transfer module is lightweight and data-efficient. By leveraging a small set of distribution control parameters, it adjusts image styles while preserving fine-grained visual details. Consequently, the generated images exhibit lighting and stylistic characteristics that closely resemble those of the target domain while effectively preserving local information.

Additionally, in Fig. 13, we provide a t-SNE visualization of the features from the new, old, and aligned data using the same feature extractor. The results show that the distribution aligning mechanism effectively reduces the domain gap between the new and old data, thereby enhancing the interaction between new and historical knowledge during lifelong learning.

3) *Feature Distribution of Different Methods:* We visualize the features of different domains in Fig. 14, comparing our DKP and DKP++ with the prototype-based method PASS and the knowledge distillation-based method LSTKC. Both PASS and LSTKC struggle in retaining sufficient knowledge for both new and old domains, primarily due to their limited ability to utilize historical knowledge effectively. In contrast, our DKP method enables clear separation between both new and old identities, demonstrating superior knowledge accumulation. This capability is primarily attributed to our anisotropic prototype generation mechanism and prototype knowledge transfer loss  $\mathcal{L}_{proto}$ , which not only facilitates the acquisition of new knowledge but also reinforces the retention of previously learned information.

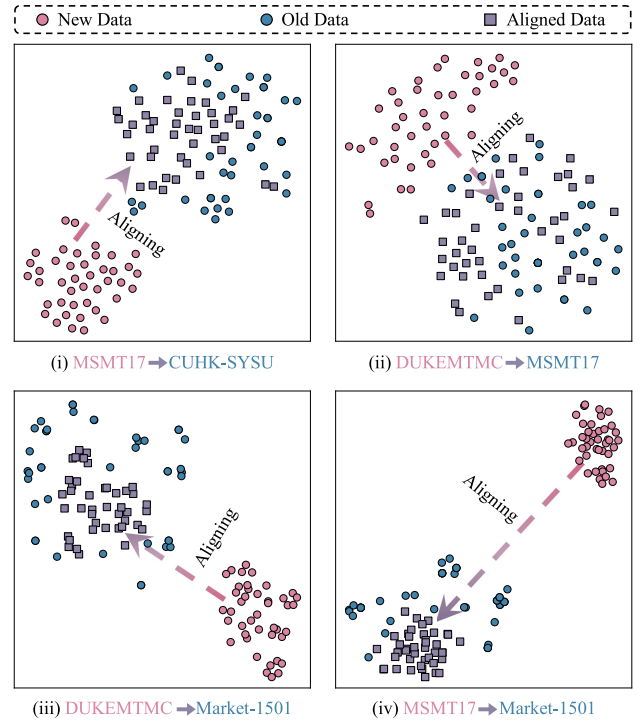


Fig. 13. Influence of distribution alignment on instance features. The feature distribution drift between new and old data is significantly reduced by distribution alignment.

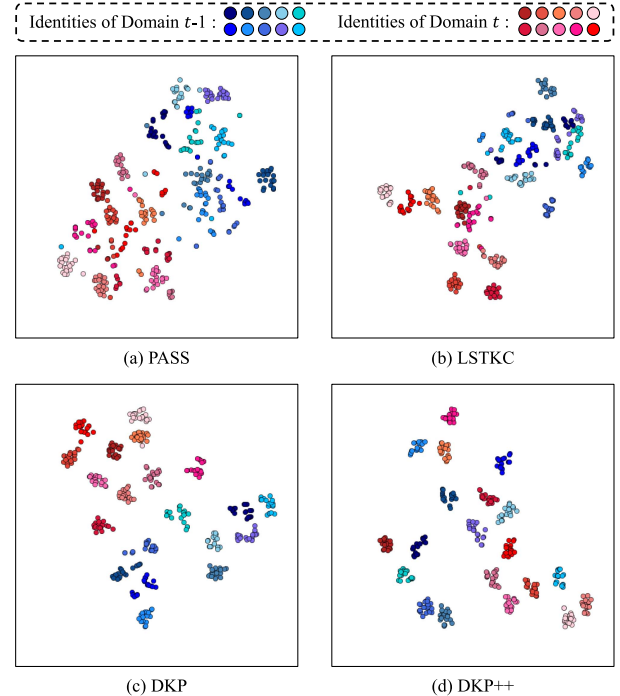


Fig. 14. t-SNE visualization of the features extracted by different methods.

Additionally, as illustrated in Fig. 14(d), our DKP++ not only distinguishes different identities but also exhibits a mix of feature distributions across various domains. In contrast, other methods tend to automatically segregate features by domain. This separation arises because the identities between datasets

do not overlap, leading LReID models to rely on apparent domain gaps, such as illumination, to differentiate identities across different domains. However, this reliance on domain gaps often involves person-irrelevant information, which can hinder the model's generalization capacity. The domain aligning scheme in DKP++ addresses this by allowing diverse domain information to be associated with the same identity. This facilitates the model to learn and retrain person-relevant knowledge instead of domain-relevant knowledge. This property not only improves the seen domain knowledge retention capacity but also enhances the cross-domain generalization capacity as verified in Fig. 8.

## VI. CONCLUSION

In this paper, we focus on the non-exemplar LReID and propose a novel method, Distribution-aware Knowledge Aligning and Prototyping (DKP++), which pioneers the integration of prototypes into the LReID setting. To encode more fine-grained knowledge within the prototypes, we introduce an Instance-level Distribution Modeling module that captures instance-specific distributional characteristics. Then, a Distribution-oriented Prototype Generation module aggregates these instance-specific distributions into identity-level prototypes. To effectively leverage the preserved historical knowledge, we further propose a Prototype-based Knowledge Transfer module that distills information from the stored prototypes into the evolving model. Furthermore, to enhance the utilization of prototypes under domain shifts, we introduce a dual-side alignment scheme. On one hand, a Distribution Aligning Module is proposed to improve coherence between new and historical knowledge by transforming new images to align with the historical image distribution using a lightweight and privacy-preserving transfer network. On the other hand, feature-level alignment constraints are applied jointly to the original and aligned data to improve knowledge acquisition and mitigate forgetting simultaneously. Extensive experiments demonstrate the superiority of our method over existing approaches. This work highlights the importance of precisely modeling and leveraging distributional information to advance lifelong learning in person re-identification.

## REFERENCES

- [1] E. Belouadah and A. Popescu, "DeeSIL: Deep-shallow incremental learning," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2018, pp. 151–157.
- [2] T. L. Hayes and C. Kanan, "Lifelong machine learning with deep streaming linear discriminant analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2020, pp. 220–221.
- [3] M. Hersche, G. Karunaratne, G. Cherubini, L. Benini, A. Sebastian, and A. Rahimi, "Constrained few-shot class-incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9057–9067.
- [4] F. Zhu, X.-Y. Zhang, C. Wang, F. Yin, and C.-L. Liu, "Prototype augmentation and self-supervision for incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5871–5880.
- [5] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3908–3916.
- [6] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2285–2294.
- [7] G. Wu and S. Gong, "Generalising without forgetting for lifelong person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2889–2897.
- [8] W. Ge et al., "Lifelong person re-identification by Pseudo task knowledge preservation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 688–696.
- [9] 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. 7897–7906.
- [10] M. Toldo and M. Oza, "Bring evanescent representations to life in lifelong class incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16732–16741.
- [11] 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.*, 2023, pp. 3295–3303.
- [12] 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.
- [13] Z. Sun and Y. Mu, "Patch-based knowledge distillation for lifelong person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 696–707.
- [14] K. Xu, X. Zou, and J. Zhou, "LSTKC: Long short-term knowledge consolidation for lifelong person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 16202–16210.
- [15] K. Xu, H. Zhang, Y. Li, Y. Peng, and J. Zhou, "Mitigate catastrophic remembering via continual knowledge purification for noisy lifelong person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 1–9.
- [16] L. Yu et al., "Semantic drift compensation for class-incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6982–6991.
- [17] B. Zhang, X. Li, Y. Ye, Z. Huang, and L. Zhang, "Prototype completion with primitive knowledge for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3754–3762.
- [18] D. Goswami, Y. Liu, B. Twardowski, and J. Van De Weijer, "FeCAM: Exploiting the heterogeneity of class distributions in exemplar-free continual learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 6582–6595.
- [19] 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. 16604–16613.
- [20] J. Shi et al., "Dual Pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 11218–11228.
- [21] 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.
- [22] 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. 17343–17353.
- [23] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 14993–15002.
- [24] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5157–5166.
- [25] X. Zhu, X. Zhu, M. Li, P. Morerio, V. Murino, and S. Gong, "Intra-camera supervised person re-identification," *Int. J. Comput. Vis.*, vol. 129, pp. 1580–1595, 2021.
- [26] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7195–7204.
- [27] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 719–728.
- [28] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3143–3152.
- [29] Y. Zou, X. Yang, Z. Yu, B. V. Kumar, and J. Kautz, "Joint disentangling and adaptation for cross-domain person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 87–104.
- [30] Q. Li, K. Xu, Y. Peng, and J. Zhou, "Exemplar-free lifelong person re-identification via prompt-guided adaptive knowledge consolidation," *Int. J. Comput. Vis.*, vol. 132, no. 11, pp. 4850–4865, 2024.
- [31] Z. Cui, Y. Peng, X. Wang, M. Zhu, and J. Zhou, "Continual vision-language retrieval via dynamic knowledge rectification," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 11704–11712.
- [32] B. Zhao, X. Xiao, G. Gan, B. Zhang, and S.-T. Xia, "Maintaining discrimination and fairness in class incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13205–13214.
- [33] Z. Huang et al., "Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14288–14297.

- [34] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICARL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5533–5542.
- [35] J. Huang et al., "Learning consistent region features for lifelong person re-identification," *Pattern Recognit.*, vol. 144, 2023, Art. no. 109837.
- [36] Z. Yang, D. Wu, W. Zhang, B. Li, and W. Wang, "Handling label uncertainty for camera incremental person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 6253–6263.
- [37] W. Shi and M. Ye, "Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 1772–1781.
- [38] K. Zhu, W. Zhai, Y. Cao, J. Luo, and Z.-J. Zha, "Self-sustaining representation expansion for non-exemplar class-incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9296–9305.
- [39] Y. Wei, J. Ye, Z. Huang, J. Zhang, and H. Shan, "Online prototype learning for online continual learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 18764–18774.
- [40] N. Asadi, M. Davari, S. Mudur, R. Aljundi, and E. Belilovsky, "Prototype-sample relation distillation: Towards replay-free continual learning," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 1093–1106.
- [41] K. Zhu, Y. Cao, W. Zhai, J. Cheng, and Z.-J. Zha, "Self-promoted prototype refinement for few-shot class-incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6801–6810.
- [42] Q. Li, Y. Peng, and J. Zhou, "FCS: Feature calibration and separation for non-exemplar class incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 28495–28504.
- [43] O. Zohar, K.-C. Wang, and S. Yeung, "PROB: Probabilistic objectness for open world object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11444–11453.
- [44] Y. Li, Y. Duan, Z. Kuang, Y. Chen, W. Zhang, and X. Li, "Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1447–1455.
- [45] T. Yu, D. Li, Y. Yang, T. M. Hospedales, and T. Xiang, "Robust person re-identification by modelling feature uncertainty," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 552–561.
- [46] Z. Zheng and Y. Yang, "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1106–1120, 2021.
- [47] F. Lu, K. Zhu, W. Zhai, K. Zheng, and Y. Cao, "Uncertainty-aware optimal transport for semantically coherent out-of-distribution detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3282–3291.
- [48] Z. Fang, R. Wang, T. Huang, and L. Jing, "Strong transferable adversarial attacks via ensembled asymptotically normal distribution learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 24841–24850.
- [49] M. Danelljan, L. V. Gool, and R. Timofte, "Probabilistic regression for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7183–7192.
- [50] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 581–588.
- [51] C. Liu, X. Chang, and Y.-D. Shen, "Unity style transfer for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6887–6896.
- [52] Y. Chong, C. Peng, J. Zhang, and S. Pan, "Style transfer for unsupervised domain-adaptive person re-identification," *Neurocomputing*, vol. 422, pp. 314–321, 2021.
- [53] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 994–1003.
- [54] N. Carlini et al., "Extracting training data from diffusion models," in *32nd USENIX Secur. Symp.*, 2023, pp. 5253–5270.
- [55] H. Piao, Y. Wu, D. Wu, and Y. Wei, "Federated continual learning via prompt-based dual knowledge transfer," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 40725–40739.
- [56] J. Wu, H. Liu, W. Shi, M. Liu, and W. Li, "Style-agnostic representation learning for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 26, pp. 2263–2275, 2023.
- [57] K. Xu, C. Jiang, P. Xiong, Y. Peng, and J. Zhou, "DASK: Distribution rehearsing via adaptive style kernel learning for exemplar-free lifelong person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 8915–8923.
- [58] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
- [59] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, pp. IV–317.
- [60] X. Xing, T. Wu, S.-C. Zhu, and Y. N. Wu, "Inducing hierarchical compositional model by sparsifying generator network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14296–14305.
- [61] A. Rényi, *Probability Theory*, North Chelmsford, MA, USA: Courier Corporation, 2007.
- [62] Y. Chen, C. Li, and Z. Lu, "Computing wasserstein-P distance between images with linear cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 519–528.
- [63] H. Alzer, "On the Cauchy-Schwarz inequality," *J. Math. Anal. Appl.*, vol. 234, no. 1, pp. 6–14, 1999.
- [64] A. Pakman et al., "Estimating the unique information of continuous variables," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 20295–20307.
- [65] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.
- [66] 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.
- [67] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," 2016, *arXiv:1604.01850*.
- [68] 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.
- [69] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian. Conf. Comput. Vis.*, Springer, 2012, pp. 31–44.
- [70] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3594–3601.
- [71] 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.
- [72] M. Hirzer, C. Beleznaï, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*. Berlin, Germany: Springer, 2011, pp. 91–102.
- [73] H. O. S. D. Branch, "Imagery library for intelligent detection systems (I-LIDS)," in *Proc. IET Conf. Crime Secur. IET*, 2006, pp. 445–448.
- [74] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *Int. J. Comput. Vis.*, vol. 90, no. 1, pp. 106–129, 2010.
- [75] H. Zhao et al., "Spindle Net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 907–915.
- [76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [77] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
- [78] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.
- [79] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1365–1374.
- [80] B. Zhao, S. Tang, D. Chen, H. Bilen, and R. Zhao, "Continual representation learning for biometric identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1197–1207.
- [81] N. Pu, Z. Zhong, N. Sebe, and M. S. Lew, "A memorizing and generalizing framework for lifelong person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13567–13585, Nov. 2023.
- [82] L. Zhang, G. Gao, and H. Zhang, "Spatial-temporal federated learning for lifelong person re-identification on distributed edges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 2, pp. 1884–1896, Feb. 2025.
- [83] D. Fu et al., "Unsupervised pre-training for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14750–14759.
- [84] D. Fu et al., "Large-scale pre-training for person re-identification with noisy labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2476–2486.
- [85] K. Zhu, H. Guo, T. Yan, Y. Zhu, J. Wang, and M. Tang, "PASS: Part-aware self-supervised pre-training for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 198–214.
- [86] J. Zuo et al., "Cross-video identity correlating for person re-identification pre-training," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 25228–25250.



**Jiahuan Zhou** received the BE degree from Tsinghua University, in 2013, and the PhD degree in the Department of Electrical Engineering & Computer Science, Northwestern University, in 2018. Currently, he is a Tenure-Track assistant professor with the Wangxuan Institute of Computer Technology, Peking University. His research interests include computer vision, deep learning, and machine learning. He has authored over 60 papers in international journals and conferences including *Nature Communications*, *Nature Synthesis*, *IEEE Transactions on Pattern Analysis and Machine*

*Intelligence*, *International Journal of Computer Vision*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Information Forensics and Security*, *CVPR*, *ICCV*, *ICML*, *ECCV*, *ICLR*, *AAAI*, *ACM MM*, and so on. He serves as an area chair for *CVPR*, *ICML*, *NeurIPS*, *ICME*, and *ICPR*, an associate editor of the *Springer Journal of Machine Vision and Applications (MVA)*, a regular reviewer member for several journals and conferences, e.g., *T-PAMI*, *IJCV*, *TIP*, *CVPR*, *ICCV*, *ECCV*, *NeurIPS*, *ICML*, *ICLR*, *AAAI*, and so on.



**Xu Zou** (Member, IEEE) received the PhD degree from the Huazhong University of Science and Technology, Wuhan, China, in 2020. He is currently a lecturer with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests include image processing, computer vision, and intelligent system.



**Kunlun Xu** received the BE and MS degrees from the Huazhong University of Science and Technology, in 2020 and 2023, respectively. He is currently working toward the PhD degree with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include computer vision, lifelong learning, and person re-identification. He has authored 13 papers in international journals and conferences, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IJCV*, *ICML*, *CVPR*, *ICCV*, *AAAI*, *ACM MM*, and so on.

He serves as a regular reviewer member for a number of journals and conferences, e.g., *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Multimedia*, *ICML*, *NeurIPS*, *CVPR*, *ICCV*, *AAAI*, *ACM MM*, and so on.



**Yuxin Peng** (Senior Member, IEEE) received the PhD degree in computer application technology from Peking University, Beijing, China, in 2003. He is the Boya distinguished professor with the Wangxuan Institute of Computer Technology, Peking University. He was a recipient of the National Science Fund for Distinguished Young Scholars of China. His research interests mainly include cross-media analysis and reasoning, image and video recognition and understanding, and computer vision. He has authored over 200 papers, including more than 100 papers in top-tier

journals and conference proceedings. He has submitted 51 patent applications and has been granted 39 of them. He led his team to win the First Place in the video semantic search evaluation of TRECVID ten times in recent years. He won the First Prize of the Beijing Science and Technology Award in 2016 (ranking first) and the First Prize of the Scientific and Technological Progress Award of the Chinese Institute of Electronics in 2020 (ranking first). He was a recipient of the Best Paper award at *MMM* 2019 and *NCIG* 2018, and serves as the associate editor of *IEEE Transactions on Multimedia*, *TCSVT*, etc.



**Fan Zhuo** is currently working toward the bachelor of engineering degree in cyberspace security with Xidian University, Xi'an, China. He is also an intern with the Wangxuan Institute of Computer Technology, Peking University. His research interests include continual learning and person re-identification.