Distribution-aware Knowledge Prototyping for Non-exemplar Lifelong Person Re-identification

Kunlun Xu[1], Xu Zou[2], Yuxin Peng[1], Jiahuan Zhou[1*]

[1] Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China

[2] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China

xkl@stu.pku.edu.cn, zoux@hust.edu.cn, {pengyuxin, jiahuanzhou}@pku.edu.cn

Abstract

Lifelong person re-identification (LReID) suffers from the catastrophic forgetting problem when learning from non-stationary data. Existing exemplar-based and knowledge distillation-based LReID methods encounter data privacy and limited acquisition capacity respectively. In this paper, we instead introduce the prototype, which is under-investigated in LReID, to better balance knowledge forgetting and acquisition. Existing prototype-based works primarily focus on the classification task, where the prototypes are set as discrete points or statistical distributions. However, they either discard the distribution information or omit instance-level diversity which are crucial fine-grained clues for LReID. To address the above problems, we propose Distribution-aware Knowledge Prototyping (DKP) where the instance-level diversity of each sample is modeled to transfer comprehensive fine-grained knowledge for prototyping and facilitating LReID learning. Specifically, an Instance-level Distribution Modeling network is proposed to capture the local diversity of each instance. Then, the Distribution-oriented Prototype Generation algorithm transforms the instance-level diversity into identity-level distributions as prototypes, which is further explored by the designed Prototype-based Knowledge Transfer module to enhance the knowledge anti-forgetting and acquisition capacity of the LReID model. Extensive experiments verify that our method achieves superior plasticity and stability balancing and outperforms existing LReID methods by 8.1%/9.1% average mAP/R@1 improvement. The code is available at https://github.com/zhoujiahuan1991/CVPR2024-DKP

1. Introduction

As a conventional task in computer vision, person re-identification (ReID) [1, 19] has achieved remarkable performance in static datasets 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 [6, 34]. Therefore, recent efforts have

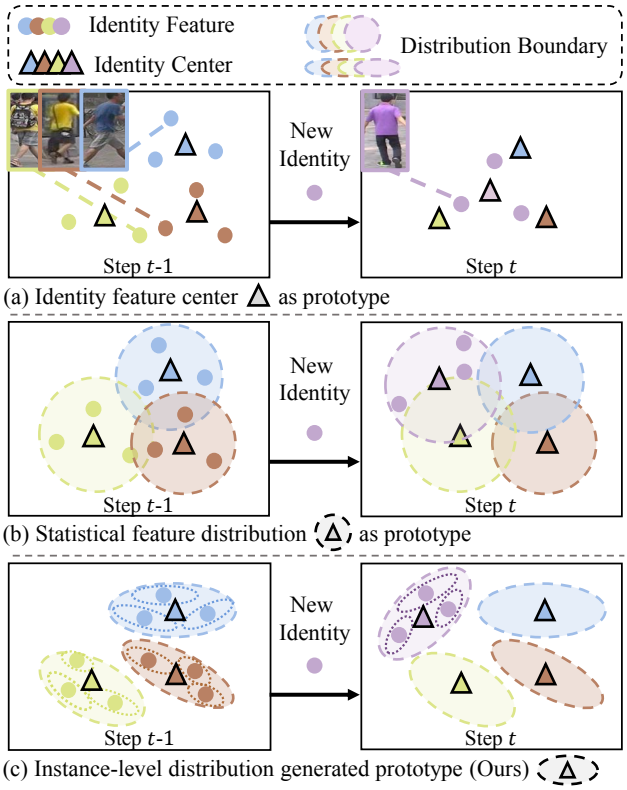


Figure 1. The comparison of different prototype generation methods. (a) Various methods [3, 8, 10] treat the prototype as a feature center point but ignore the important distribution information. (b) Some approaches [46] calculate the prototype as a statistical feature distribution but suffer from inaccurate 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.

*Corresponding author

shifted towards addressing a more challenging and practical lifelong person re-identification (LReID) problem [24, 34], where the training data from different scenarios come continually and the trained model can not only acquire new information but also preserve already learned old knowledge.

As well-known in existing lifelong learning tasks [30], 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, as a fine-grained task, the data distribution for the same person in LReID can significantly vary, stemming from temporal and environmental changes. 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 [6, 34, 36], but inevitably raise severe privacy concerns and computational costs [25]. As an alternative, various rehearsal-free methods [24–26, 29] 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 learning of the new data.

Recently, a few works aim to explore prototypes to mitigate catastrophic forgetting in other lifelong learning tasks, *e.g.*, class incremental learning (CIL) [30]. These prototype-based CIL methods [8, 10, 30, 37] usually calculate or learn the feature center of a class as the prototype in Fig. 1 (a). Nevertheless, solely using a feature point as the prototype ignores the intra-class diversity information, resulting in serious forgetting caused by insufficient distribution knowledge of historical data [46]. Thus, several methods [39, 46] consider calculating a statistic 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 individual characteristics. As mentioned above, in LReID, samples from the same person can show significant differences because of temporal and environmental changes. Thus, their obtained prototypes 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.

In this paper, we propose a novel non-exemplar LReID method named Distribution-aware Knowledge Prototyping (DKP), that readily models the instance-level distribution of each sample to generate a more informative person-specific prototype as shown in Fig. 1 (c), thereby transferring the useful distribution knowledge to the new model and achieving better anti-forgetting capacity. To this end, an Instance-level Distribution Modeling (IDM) network is designed to estimate the distribution for each input instance, which is accomplished based on the sampling strategy and 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 predicted 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 discriminant of new identity features with the aid of historical distributions, which mitigates the forgetting of the old knowledge and guarantees new knowledge acquisition. In summary, our contributions are three-fold: (1) A novel non-exemplar LReID method is proposed that models the instance-level distribution for each input sample to achieve distribution-aware prototyping. (2) An effective knowledge transfer scheme is designed to fully explore the obtained distribution-aware prototypes for enhancing the discriminant across datasets and consolidating the learned knowledge. (3) Extensive experiments on various datasets and settings have verified the superiority of our method against the state-of-the-art LReID approaches.

2. Related work

2.1. Lifelong Person Re-Identification

Lifelong person re-identification (LReID) aims to improve the person matching capacity of the model by learning from non-stationary data [24, 34], where catastrophic forgetting is the key challenge [30, 41]. Existing LReID works could be categorized into two branches, rehearsal-based and knowledge distillation-based. The former ones [6, 15, 34, 36] aim to mitigate forgetting by storing the exemplar images of previous steps and replaying them when learning the new data. However, storing human images is often impractical in real scenarios due to privacy. The knowledge distillation-based approaches [14, 24–26, 29] try to preserve the old knowledge by constraining the output consistency between the old and new models. Though promising anti-forgetting capacity has been exhibited, their acquisition capacity of the new data is limited due to such strict constraints hindering model plasticity. Instead, in this paper, we investigate the non-exemplar LReID scenario by leveraging prototypes of different persons for anti-forgetting.

2.2. Prototype-based Class Incremental Learning

Recently, various prototype-based class incremental learning (CIL) methods are proposed to continually learn new classes without preserving any historical exemplars [28, 37, 46, 48]. Some methods treat the prototype of each class as a

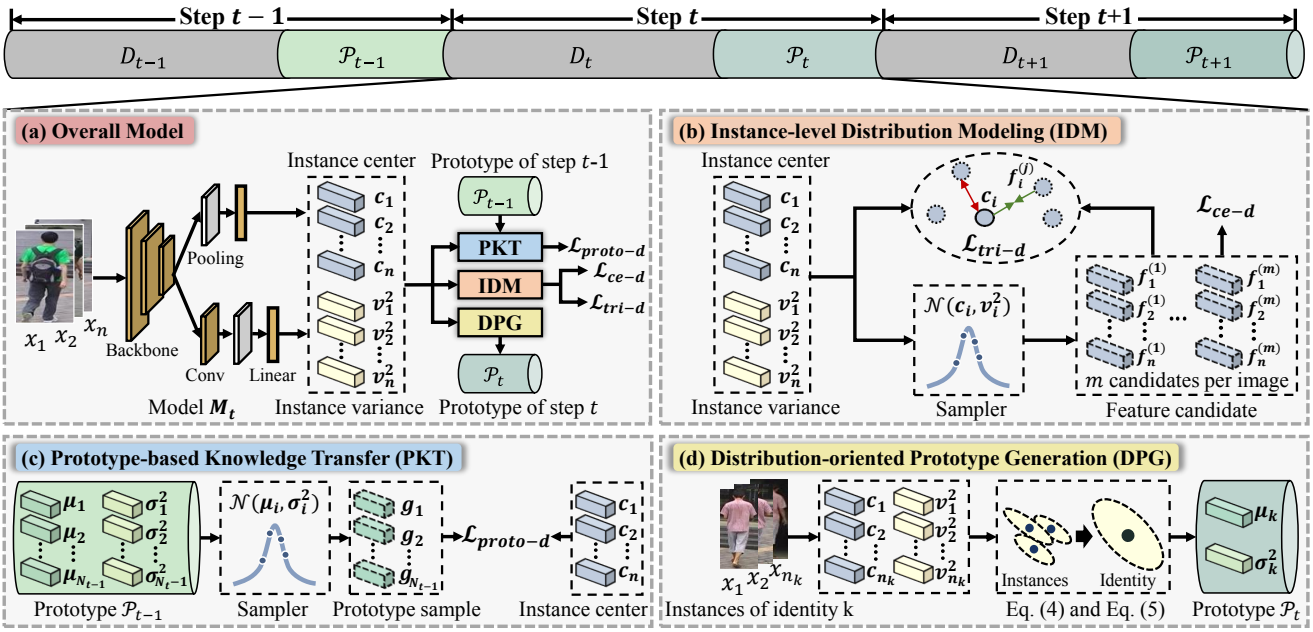


Figure 2. The pipeline of our proposed Distribution-aware Knowledge Prototyping (DKP) model. (a) Our model is built upon a dual-branch convolution network where the branches predict the instance-specific distribution center and variance, respectively. Furthermore, three novel modules are designed to accomplish the learning procedure. (b) The Instance-level Distribution Modeling (IDM) module incorporates the random sampling and the ReID losses $\mathcal{L}_{trip-d}$, $\mathcal{L}_{ce-d}$ to facilitate the distribution learning of each instance. (c) The Prototype-based Knowledge Transfer (PKT) module utilizes the previous step prototypes $\mathcal{P}_{t-1}$ to mitigate forgetting. (d) The Distribution-oriented Prototype Generation (DPG) transforms the learned insistence-level center and variance into prototypes of the current step $\mathcal{P}_t$.

learnable embedding vector [2, 33, 47], and others calculate the mean feature of all samples from the same class as the prototype [10, 28, 46]. When learning the new classes, the obtained prototypes are utilized to represent the knowledge of old classes for classifier training [28]. Considering using a single feature center point will inevitably result in the lack of informative distribution of data, the latest methods [46] proposed to simultaneously calculate the mean feature vector and its variance from all samples to depict the distribution information of a class. However, such a strategy simply assumes each sample has the same impact on the estimated distribution regardless of the intra-class diversity of samples. Therefore, the obtained prototype exhibits a distribution drift, thereby misleading the sequential learning steps.

2.3. Distribution Learning

In computer vision, the information of distribution is crucial to describe the inherent probabilistic knowledge of data [20, 49]. Existing distribution learning methods focus on modeling the data uncertainty to handle the out-of-distribution data [23, 38, 45]. Specifically, [20] estimated the uncertainty of labels in domain adaptive semantic segmentation to recognize and rectify the noise labels. Similarly, [38] adopted distribution learning to alleviate the negative impacts of label noises and outliers on model training. In this paper, we propose to model the instance-level distribution of each sample in LReID, based on which the fine-grained instance-specific information could be mined and integrated into the obtained distribution-aware prototypes to mitigate catastrophic forgetting of historical knowledge.

3. The Proposed DKP Method

3.1. Problem Formulation

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 previous t-1 datasets are unavailable. Given a dataset D_t with N_t identities, our method obtains a prototype set $\mathcal{P}_t = \{\boldsymbol{p}_i\}_{i=1}^{N_t}$ for all identities. For each instance, our method learns a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{c}, \boldsymbol{v}^2)$, where $\boldsymbol{c} \in \mathbb{R}^d$ is the feature center vector of dimension d, and $\boldsymbol{v}^2 \in \mathbb{R}^d$ is the diagonal element vector derived from the diagonal covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$.

3.2. The Overall Pipeline of DKP

At the t-th learning step, the overall pipeline of our proposed DKP model, as depicted in Fig. 2 (a), is based on a dual-branch convolutional network $\boldsymbol{M}_t$. Given a set of n input images $\{x_i\}_{i=1}^{n}$ from D_t, a backbone network is utilized to extract image features. Then, one branch consisting of a pooling layer followed by a linear layer is responsible for predicting the distribution centers of each input instance, denoted as $\mathcal{C} = \{\boldsymbol{c}_i\}_{i=1}^{n}$. Meanwhile, the other

branch which comprises a convolution layer, followed by a pooling layer and a linear layer, is adopted to estimate the distribution variances of the instances $\mathcal{V} = \{\boldsymbol{v}_i^2\}_{i=1}^n$. To enhance the model capability of knowledge acquisition and anti-forgetting, three innovative modules are accordingly designed to leverage the prototypes from the previous step $\mathcal{P}_{t-1}$, as well as generate prototypes of the current step $\mathcal{P}_t$.

3.3. Instance-level Distribution Modeling

In this section, we aim to guide the model to learn the instance-level distribution which contains fine-grained data knowledge as illustrated in Fig. 2 (b). Specifically, once the instance centers $\mathcal{C} = \{\boldsymbol{c}_i\}_{i=1}^n$ and variances $\mathcal{V} = \{\boldsymbol{v}_i^2\}_{i=1}^n$ are predicted by the model, for each instance x_i, we adopt a Gaussian sampler parameterized as $\mathcal{N}(\boldsymbol{c}_i, \boldsymbol{v}_i^2)$ to sample m feature candidates $\{\boldsymbol{f}_i^{(j)}\}_{j=1}^m$, where each $\boldsymbol{f}_i^{(j)}$ is assigned the same identity label as $\boldsymbol{c}_i$. To ensure identity consistency between $\{\boldsymbol{f}_i^{(j)}\}_{j=1}^m$ and $\boldsymbol{c}_i$, a distribution-aware cross-entropy loss $\mathcal{L}_{ce-d}$ is calculated as below:

$$\mathcal{L}_{ce-d} = \frac{1}{m+1}[y_i \log \boldsymbol{\rho}(\mathbf{W}_t \boldsymbol{c}_i) + \sum_{j=1}^{m} y_i \log \boldsymbol{\rho}(\mathbf{W}_t \boldsymbol{f}_i^{(j)})], \quad (1)$$

where y_i is the identity label of image x_i, $\boldsymbol{\rho}$ represents the softmax function, and $\mathbf{W}_t$ denotes the linear projection parameter to map the features to logits.

Moreover, we further extend the widely used triplet loss in existing LReID models [29, 36] to a distribution-oriented version, denoted as $\mathcal{L}_{tri-d}$. Given n images $\{x_i\}_{i=1}^n$ and their sampling set $\mathcal{F}_s = \{(\boldsymbol{f}_i^{(j)}, y_i) | 1 \leq i \leq n, 1 \leq j \leq m\}$, the $\mathcal{L}_{tri-d}$ for an instance x_i is calculated by:

$$\mathcal{L}_{tri-d} = \log\left(1 + \exp(\| \boldsymbol{c}_i - \boldsymbol{f}'_p \|_2^2 - \| \boldsymbol{c}_i - \boldsymbol{f}'_n \|_2^2)\right), \quad (2)$$

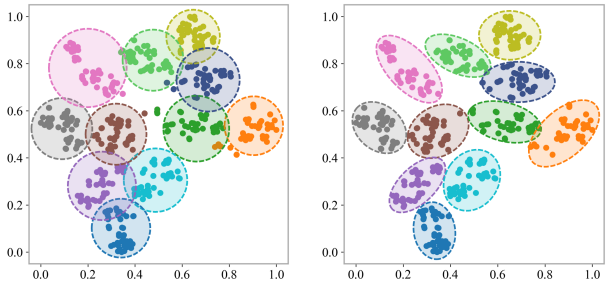
where $\boldsymbol{c}_i$, $\boldsymbol{f}'_p$, and $\boldsymbol{f}'_n$ indicate the anchor point, positive point, and negative point respectively. Specifically, $\boldsymbol{f}'_p$ and $\boldsymbol{f}'_n$ is obtained by:

$$\begin{cases} \boldsymbol{f}'_p = \arg \max\limits_{(\boldsymbol{f}', y') \in \mathcal{F}_s, y' = y_i} \| \boldsymbol{c}_i - \boldsymbol{f}' \|_2^2 \\ \boldsymbol{f}'_n = \arg \min\limits_{(\boldsymbol{f}', y') \in \mathcal{F}_s, y' \neq y_i} \| \boldsymbol{c}_i - \boldsymbol{f}' \|_2^2 \end{cases}, \quad (3)$$

where $(\boldsymbol{f}', y')$ is the feature candidate and identity label pair. Based on $\mathcal{L}_{ce-d}$ and $\mathcal{L}_{tri-d}$, the instances of the same identity can be pushed together and the ones of different identities would be pulled away which promotes discriminative knowledge learning. Besides, when an instance is away from its corresponding identity center, it tends to learn a larger $\boldsymbol{v}_i^2$ to generate samples closer to other instances of the same identity [5, 38].

3.4. Distribution-oriented Prototype Generation

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 $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$ where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and $\boldsymbol{\sigma}_k^2 \in \mathbb{R}^d$ 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 $\boldsymbol{\mu}_k$ could be obtained by:



(a) PASS (b) Ours

Figure 3. Visualization of our instance-level distribution-aware prototype and statistical feature distribution-based prototype [46].

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{c}_i, \quad (4)$$

where $\boldsymbol{c}_i$ is the learned enter of the instance x_i. Besides, the identity distribution variance $\boldsymbol{\sigma}_k^2$ could be obtained by:

$$\begin{aligned} \boldsymbol{\sigma}_k^2 &= \int \boldsymbol{x}^2 \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) d\boldsymbol{x} - (\int \boldsymbol{x} \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) d\boldsymbol{x})^2 \\ &= \frac{1}{n_k} \sum_{i=1}^{n_k} (\boldsymbol{c}_i^2 + \boldsymbol{v}_i^2) - (\frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{c}_i)^2, \end{aligned} \quad (5)$$

where

$$\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) = \frac{e^{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)}}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} \quad (6)$$

is the distribution function denoted by $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$ and $\boldsymbol{\Sigma}_k$ is a diagonal matrix whose diagonal elements are $\boldsymbol{\sigma}_k^2$. Similarly, $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{c}_i, \boldsymbol{v}_i)$ is the distribution functions denoted by $\mathcal{N}(\boldsymbol{c}_i, \boldsymbol{v}_i^2)$. More derivation of Eq. (4) and Eq. (5) is provided in the Supplementary Material. Compared to the statistical feature distribution-based prototypes [46] that only utilize $\boldsymbol{c}_i$ to obtain identity-level divergence, our proposed distribution-oriented prototype simultaneously considers $\boldsymbol{c}_i$ and $\boldsymbol{v}_i$ when calculating $\boldsymbol{\sigma}_k^2$, making our prototype more informative. To be noted, $\boldsymbol{v}_i^2$ is important in describing the precise distribution of an identity. As illustrated in Fig. 3, our predicted distribution could better depict the area of identities with fewer outliers and more discriminative boundaries between identities.

In this paper, $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k^2$ are calculated after the completion of the t-th training step, and $\mathcal{P}_t = \{(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)\}_{k=1}^{N_t}$ is preserved for the next step.

	Method	Market1501		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
	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
CIL	LwF [21]	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 [31]	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
	PRAKA* [28]	37.4	61.3	69.3	71.8	35.4	55.0	10.7	27.2	**54.0**	**55.6**	41.3	54.2	47.7	41.6
	PRD* [2]	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
LReID	CRL [42]	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 [24]	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† [24]	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 [29]	**68.5**	**85.7**	75.6	78.6	33.8	50.4	6.5	17.0	34.1	36.8	43.7	53.7	49.1	45.4
	MEGE [26]	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
	DKP(Ours)	60.3	80.6	**83.6**	**85.4**	**51.6**	**68.4**	**19.7**	**41.8**	43.6	44.2	**51.8**	**64.1**	**59.2**	**51.6**

Table 1. Training Order-1: Market-1501 → CUHK-SYSU → DukeMTMC-reID → MSMT17-V2→ CUHK03. $*$ denotes the results are reproduced by the released official code. † denotes the results reported by [29].

	Method	DukeMTMC		MSMT17		Market1501		CUHK-SYSU		CUHK03		Seen-Avg		Unseen-Avg	
		mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1
	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
CIL	LwF [21]	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 [31]	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
	PRAKA* [28]	31.2	48.7	6.6	19.1	47.8	69.8	70.4	73.0	**54.9**	**56.6**	42.2	53.4	48.4	41.1
	PRD* [2]	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
LReID	CRL [42]	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 [24]	32.5	49.7	-	-	-	-	-	-	-	-	-	-	40.8	37.2
	AKA† [24]	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 [29]	**58.3**	**74.1**	6.4	17.4	43.2	67.4	74.5	76.9	33.7	34.8	43.2	54.1	48.6	44.1
	MEGE [26]	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
	DKP(Ours)	53.4	70.5	**14.5**	**33.3**	**60.6**	**81.0**	**83.0**	**84.9**	45.0	46.1	**51.3**	**63.2**	**59.0**	**51.6**

Table 2. Training Order-2: DukeMTMC-reID → MSMT17-V2→ Market-1501 → CUHK-SYSU → CUHK03. $*$ denotes the results are reproduced by the released official code. † denotes the results reported by [29]. '-' denotes the result was not reported in the original paper.

3.5. Prototype-based Knowledge Transfer

In our method, the prototypes from the previous step describe the old identity distributions in the feature space. When new data is introduced, the extracted new features should be distinguishable from the old distributions. To achieve this, we propose a Prototype Knowledge Transfer scheme: Firstly, as illustrated in Fig. 2 (c), given a set of prototypes $\mathcal{P}_{t-1} = \{(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)\}_{i=1}^{N_{t-1}}$, we sample N_{t-1} prototype features $\mathcal{F}_p = \{\boldsymbol{g}_i\}_{i=1}^{N_{t-1}}$ according to the distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$, where $\mathcal{F}_p$ is formed as a matrix $\mathbf{F}_p \in \mathbb{R}^{N_{t-1} \times d}$. 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:

$$\mathbf{C}_p = \boldsymbol{\rho}(\mathbf{F}_c \mathbf{F}_p^\top / \lambda_1), \tag{7}$$

where the softmax function $\boldsymbol{\rho}$ is applied row-wisely and λ_1 is the temperature parameter [12] to scale the matrix values. Note that each row of $\mathbf{C}_p$ encodes the relative distance between x_i and all prototypes. Then, to ensure the discriminant between the new and old data, and additionally guide the new model learning with old knowledge, we co-optimize the inter-instance affinity with the proposed Prototype Knowledge Transfer loss which is calculated by:

$$\mathcal{L}_{proto-d} = \mathcal{L}_{KL}(\boldsymbol{\rho}(\mathbf{C}_p \mathbf{C}_p^\top / \lambda_2) || \boldsymbol{\rho}(\mathbf{F}_c \mathbf{F}_c^\top / \lambda_2)), \tag{8}$$

where $\mathcal{L}_{KL}$ is the Kullback Leibler divergence [11] and λ_2 is another temperature parameter. $\boldsymbol{\rho}(\mathbf{C}_p \mathbf{C}_p^\top)$ and $\boldsymbol{\rho}(\mathbf{F}_c \mathbf{F}_c^\top)$ 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 Eq. (8) guarantees the acquisition of new data and mitigates forgetting of old knowledge.

3.6. Training and Inference

During training, we follow the procedure illustrated in Fig. 2 and the overall loss is:

$$\mathcal{L} = \mathcal{L}_{ce-d} + \alpha \mathcal{L}_{tri-d} + \beta \mathcal{L}_{proto-d}, \tag{9}$$

where α and β are hyperparameters to balance the loss weights. In this paper, we set $\alpha = 1.5$ and $\beta = 0.1$ re-

spectively. At the end of the t-th learning step, to further blend the knowledge, we fuse the learned model $\boldsymbol{M}_t$ and the old model $\boldsymbol{M}_{t-1}$ as post-processing

$$\boldsymbol{M}_t \leftarrow (\boldsymbol{M}_t + \boldsymbol{M}_{t-1})/2. \tag{10}$$

During testing, we use the predicted feature center $\boldsymbol{c}$ of input image x as shown in Fig. 2 (a) for person matching.

4. Experiments

4.1. Experimental Settings

Benchmarks. We conduct the experiments on the LReID benchmark [24] which comprises a total of twelve ReID datasets. Among them, five datasets (Market1501 [44], DukeMTMC-reID [27], CUHK-SYSU [35], MSMT17-V2 [32], and CUHK03 [18]) are used as seen domains for lifelong training and testing. To verify the performance consistency of the models, two different dataset orders are adopted to form various lifelong learning scenarios [1] [2]. Additionally, the other seven datasets (CUHK01 [17], CUHK02 [16], VIPeR [7], PRID [13], i-LIDS [4], GRID [22], and SenseReID [43]) are tested as the unseen domains to show the generalization capacity of the models.
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. Besides, the average mAP and average R@1 on all seen and unseen domains are calculated to compare the overall lifelong learning and generalization capacity of the models respectively.
Implementation Details. For a fair comparison with existing LReID methods [24, 29], we utilize the ResNet50 architecture as our backbone. For both training orders, we train the first dataset for 80 epochs, and the subsequent datasets for 60 epochs each. A mini-batch size of 128 is adopted, where 32 identities are sampled with 4 images for each identity during training. For model optimization, the SGD optimizer with a learning rate of 0.008 and a weight decay of 0.0001 is used. Furthermore, the temperature parameter λ_1 and λ_2 are set as 0.1.

4.2. The Compared Methods

To extensively evaluate our method, various state-of-the-art non-exemplar LReID approaches including CRL [42], AKA [24], PatchKD [29], MEGE [26] are compared. Additionally, the latest class incremental learning (CIL) methods, LwF [21], SPD [40], PRAKA [28], and PRD [2], 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 [9] to align 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. Besides, we also implement *Finetune* that trains on the datasets step by step without anti-forgetting designs.

[1]Training Order-1: Market1501 → CUHK-SYSU → DukeMTMC → MSMT17 → CUHK03.

[2]Training Order-2: DukeMTMC → MSMT17 → Market1501 → CUHK-SYSU → CUHK03.

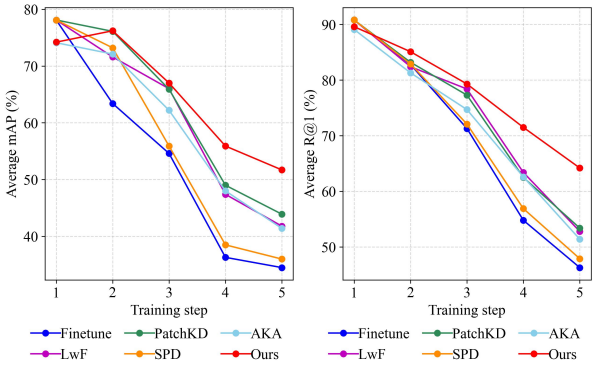


Figure 4. Performance tendency on seen domains. After each training step, the model is evaluated on the already-seen domains.

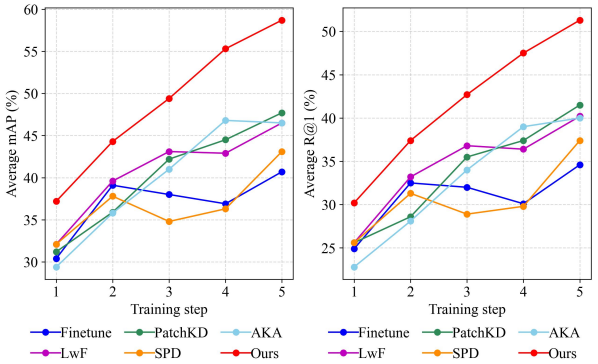


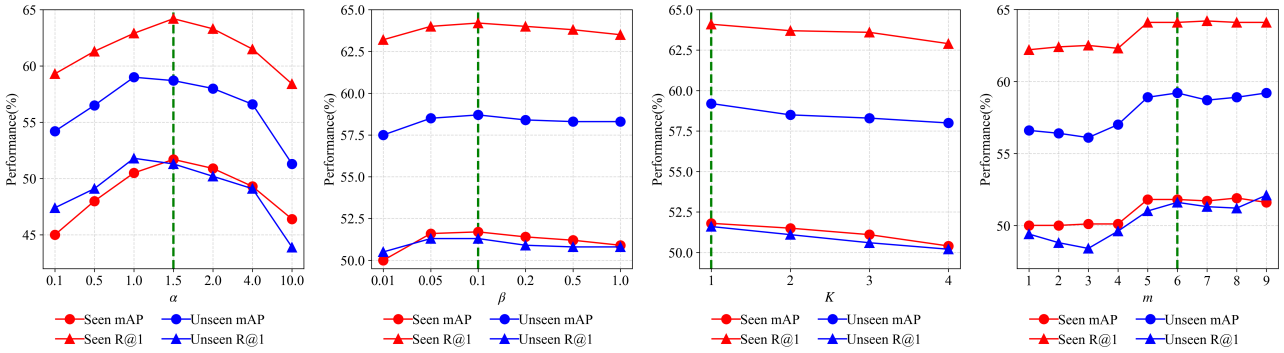
Figure 5. Performance tendency on unseen domains. After each training step, the performance of all unseen domains is evaluated.

4.3. 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 Tab. 1 and Tab. 2, corresponding to Training Order-1 and Training Order-2 respectively. The best and second best results are marked in **Bold** and underlined.

Compared to LReID Methods: As reported in Tab. 1 and Tab. 2, our DKP outperforms all existing LReID models significantly. Compared to the second-best player PatchKD, our method achieves an improvement of **8.1%/10.4%** and **8.1%/9.1%** on the average mAP/R@1 performance for seen domains. Notably, PatchKD performs better than ours on the initial dataset because it employs a strict knowledge distillation constraint that enforces output consistency on the new and old models, thereby enhancing the model's anti-forgetting capacity. However, the learnability of PatchKD is severely limited to new data, resulting in inferior performance on subsequent datasets.

Compared to CIL Methods: Our DKP outperforms the CIL methods on the Market1501, CUHK-SYSU,



(a) The weight of $\mathcal{L}_{tri-d}$ (b) The weight of $\mathcal{L}_{proto-d}$ (c) Stored prototype stages (d) Feature candidates number

Figure 6. Ablation studies on hyperparameters. The values marked by the dashed lines are adopted by our proposed method.

Baseline	IDM	DPG&PKT	Seen-Avg mAP	Seen-Avg R@1	Unseen-Avg mAP	Unseen-Avg R@1
✓			44.0	53.4	46.8	39.8
✓	✓		46.1	60.0	51.6	45.5
✓		✓	50.0	62.2	57.7	49.9
✓	✓	✓	**51.8**	**64.1**	**59.2**	**51.6**

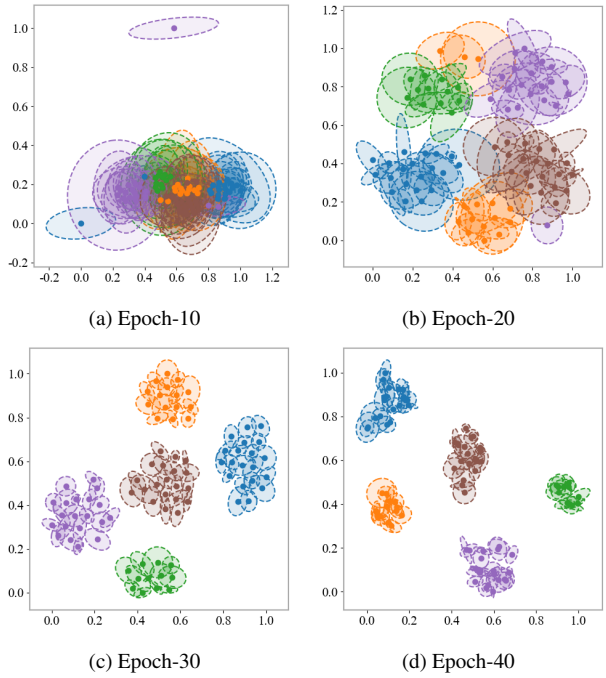
Table 3. Ablation study of different components.

DukeMTMC, and MSMT17 datasets, exhibiting an average mAP/R@1 improvement of 10.5%/9.9% and 9.1%/9.8% across both training orders for seen domains. We observe that although the prototype-based model PRAKA excels our DKP on the last domain CUHK03, our model achieves remarkably higher performance than PRAKA in the early-stage domains and superior average performance on both seen and unseen domains, owing to the knowledgeable information encoded by our distribution-aware prototypes, which mitigates the forgetting caused by domain shifts.

Seen Domain Performance Tendency. Fig. 4 demonstrates the anti-forgetting capacity of different models. Compared to other methods, our DKP initially obtains slightly inferior performance. This disparity arises mainly because our proposed instance-level distribution learning component partially slows down the convergence. With the increase in training steps, our model exhibits superior average performance. These results indicate that our method excels in the long-term consolidation of knowledge.

4.4. Unseen-Domain Generalization Evaluation

The average performance on unseen domains is shown in the last two columns of Tab. 1 and Tab. 2. Our method demonstrates superior generalization capabilities compared to state-of-the-art CIL models, exhibiting an average mAP/R@1 improvement of 11.5%/9.0% and 10.6%/10.5% across both training orders. Furthermore, our model also significantly outperforms the LReID models by a margin of 10.1%/6.2% and 10.4%/7.5% average mAP/R@1 improvement. These results substantiate that our model effectively consolidates more generalizable knowledge.



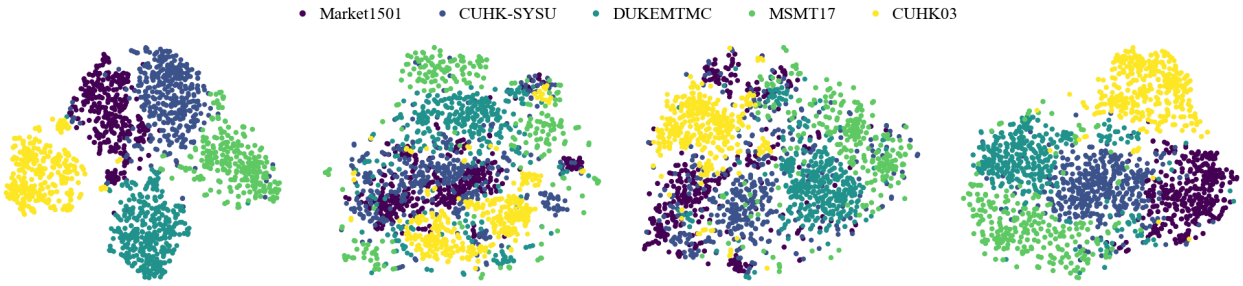
(a) Epoch-10 (b) Epoch-20 (c) Epoch-30 (d) Epoch-40

Figure 7. The visualization of the learned distributions under different training epochs.

Generalization Curves. We further analyze the average performance on the unseen domains along the lifelong training steps, as depicted in Fig. 5. The results show that our DKP could preserve more generalizable knowledge compared to existing methods. Additionally, our method exhibits faster performance growth across the training steps, further highlighting its superior generalization capability.

4.5. Ablation Studies

Influence of Different Components. As shown in Tab. 3, the *Baseline* model is the framework excluding our IDM, DPG and PKT. We then add these modules to evaluate their impacts. Because the DPG module does not independently influence the model training, we integrate it as the front module of PKT, formulating the combined module as DPG&PKT. As we can see, the utilization of both



(a) Joint-Train (b) Finetune (c) PatchKD (d) Ours

Figure 8. The t-SNE visualization results of the learned features on five seen datasets.

the IDM and DPG&PKT modules consistently leads to performance improvements on both seen and unseen domains. The instance-level distribution modeling in IDM guides the model to learn more discriminative features. Additionally, DPG&PKT facilitates the transfer of distribution knowledge encoded in the prototypes to the new model that effectively mitigates forgetting. When both two modules are used together, the performance is further improved because the instance distributions modeled by IDM promote the knowledge transfer capacity of DPG&PKT.

Influence of Hyperparameters. We analyze the effects of weights for $\mathcal{L}_{tri-d}$ and $\mathcal{L}_{proto-d}$ in Fig. 6 (a) and (b). The results indicate that the model performance varies significantly as α changes. This is because $\mathcal{L}_{tri-d}$ is powerful in guiding the model to learn discriminative knowledge. However, an excessively large value of $\mathcal{L}_{tri-d}$ can lead to optimization instability. In contrast, $\mathcal{L}_{proto-d}$ has a more stable influence on the model since it imposes a looser constraint. In this paper, we set $\alpha = 1.5$ and $\beta = 0.1$ by default according to the results in Fig. 6 (a) and Fig. 6 (b). In Fig. 6 (c), we examine the impact of storing the prototypes from previous K stages during lifelong training. The results show a declining tendency as K increases. This is due to the prototypes from older steps becoming outdated as the model evolves. Thus, we only utilize the prototypes of the previous step in this paper. Lastly, we investigated the influence of the sampled feature candidate number m in Fig. 6 (d). The findings reveal that a higher sampling number leads to better performance, as it enables more accurate modeling of the instance distribution. For the sake of performance and effectiveness, we set $m = 6$.

4.6. Visualization Results

To illustrate the distribution modeling process of our DKP, the t-SNE visualization in Fig. 7 shows the learned centers and variances of training samples at different epochs. At the initial epochs, the features exhibit a dispersed pattern and the learned variance is large. As the epoch number increases, the features become well-clustered and the learned variances are smaller. Simultaneously, the discriminative boundaries between different clusters are obtained, indicating that different individuals are better distinguished.

Besides, we also visualize the features of different datasets shown in Fig. 8 in comparison with Joint-Train, Finetune, and PatchKD. Joint-Train represents all the data is available at once and different datasets are automatically learned to be separate. Due to the non-stationary data stream, both Finetune and PatchKD struggle to accumulate enough knowledge to separate the datasets. However, owing to the effective knowledge accumulation capacity of our distribution-aware knowledge prototyping model, each dataset can be discriminatively separated. This can be attributed to our prototype knowledge transfer loss $\mathcal{L}_{proto-d}$ which facilitates both cross-step discriminant enhancement and consolidation of historical knowledge.

4.7. Discussion and Future Work

Our proposed method outperforms existing methods on the average performance over all datasets, but it still falls short compared to PatchKD and PRAKA on the first and last datasets, respectively, indicating potential space for performance improvement. Additionally, the prototypes from earlier steps do not contribute to knowledge transfer in our method due to their outdated nature, while they still contain valuable distribution information to some extent which should be investigated in the future.

5. Conclusion

In this paper, we focus on the non-exemplar LReID and propose a prototype-based method Distribution-aware Knowledge Prototyping (DKP). To encode more fine-grained knowledge into the prototypes, we propose Instance-level Distribution Modeling and Distribution-oriented Prototype Generation modules to capture the instance-level distribution and generate distribution-aware prototypes. Besides, a Prototype-based Knowledge Transfer module is developed to consolidate the old knowledge from the prototypes into the new model. Extensive experimental results show DKP outperforms existing LReID models by a large margin. To our knowledge, DKP is a pioneer LReID work that adapts the prototype to model distribution knowledge.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (62376011, 61925201, 62132001).

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916. IEEE, 2015. 1

[2] Nader Asadi, MohammadReza Davari, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Prototype-sample relation distillation: towards replay-free continual learning. In *ICML*, pages 1093–1106. PMLR, 2023. 3, 5, 6

[3] Eden Belouadah and Adrian Popescu. Deesil: Deep-shallow incremental learning. In *ECCVW*, pages 0–0. Springer, 2018. 1

[4] Home Office Scientific Development Branch. Imagery library for intelligent detection systems (i-lids). In *2006 IET Conference on Crime and Security*, pages 445–448. IET, 2006. 6

[5] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *CVPR*, pages 5710–5719, 2020. 4

[6] Wenhang Ge, Junlong Du, Ancong Wu, Yuqiao Xian, Ke Yan, Feiyue Huang, and Wei-Shi Zheng. Lifelong person re-identification by pseudo task knowledge preservation. In *AAAI*, pages 688–696, 2022. 1, 2

[7] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. Springer, 2008. 6

[8] Tyler L Hayes and Christopher Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *CVPRW*, pages 220–221. IEEE, 2020. 1, 2

[9] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 14993–15002. IEEE, 2021. 6

[10] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *CVPR*, pages 9057–9067. IEEE, 2022. 1, 2, 3

[11] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *ICASSP*, pages IV–317. IEEE, 2007. 5

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[13] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. Springer, 2011. 6

[14] Jinze Huang, Xiaohan Yu, Dong An, Yaoguang Wei, Xiao Bai, Jin Zheng, Chen Wang, and Jun Zhou. Learning consistent region features for lifelong person re-identification. *PR*, 144:109837, 2023. 2

[15] Zhipeng Huang, Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, and Zheng-jun Zha. Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation. In *CVPR*, pages 14288–14297. IEEE, 2022. 2

[16] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601. IEEE, 2013. 6

[17] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, pages 31–44. Springer, 2012. 6

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159. IEEE, 2014. 6

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294. IEEE, 2018. 1

[20] Yi Li, Yiqun Duan, Zhanghui Kuang, Yimin Chen, Wayne Zhang, and Xiaomeng Li. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In *AAAI*, pages 1447–1455, 2022. 3

[21] Zhizhong Li and Derek Hoiem. Learning without forgetting. *PAMI*, 40(12):2935–2947, 2017. 5, 6

[22] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *IJCV*, 90(1):106–129, 2010. 6

[23] Fan Lu, Kai Zhu, Wei Zhai, Kecheng Zheng, and Yang Cao. Uncertainty-aware optimal transport for semantically coherent out-of-distribution detection. In *CVPR*, pages 3282–3291. IEEE, 2023. 3

[24] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. Lifelong person re-identification via adaptive knowledge accumulation. In *CVPR*, pages 7897–7906. IEEE, 2021. 2, 5, 6

[25] Nan Pu, Yu Liu, Wei Chen, Erwin M Bakker, and Michael S Lew. Meta reconciliation normalization for lifelong person re-identification. In *ACMM*, pages 541–549, 2022. 2

[26] Nan Pu, Zhun Zhong, Nicu Sebe, and Michael S Lew. A memorizing and generalizing framework for lifelong person re-identification. *PAMI*, 2023. 2, 5, 6

[27] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35. Springer, 2016. 6

[28] Wuxuan Shi and Mang Ye. Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning. In *ICCV*, pages 1772–1781. IEEE, 2023. 2, 3, 5, 6

[29] Zhicheng Sun and Yadong Mu. Patch-based knowledge distillation for lifelong person re-identification. In *ACMMM*, pages 696–707, 2022. 2, 4, 5, 6

[30] Marco Toldo and Mete Ozay. Bring evanescent representations to life in lifelong class incremental learning. In *CVPR*, pages 16732–16741. IEEE, 2022. 2

[31] Fred Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, pages 1365–1374. IEEE, 2019. 5

[32] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88. IEEE, 2018. 6

[33] Yujie Wei, Jiaxin Ye, Zhizhong Huang, Junping Zhang, and Hongming Shan. Online prototype learning for online con-

tinual learning. In *ICCV*, pages 18764–18774. IEEE, 2023. 3

[34] Guile Wu and Shaogang Gong. Generalising without forgetting for lifelong person re-identification. In *AAAI*, pages 2889–2897, 2021. 1, 2

[35] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv:1604.01850*, 2(2):4, 2016. 6

[36] Chunlin Yu, Ye Shi, Zimo Liu, Shenghua Gao, and Jingya Wang. Lifelong person re-identification via knowledge refreshing and consolidation. In *AAAI*, pages 3295–3303, 2023. 2, 4

[37] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *CVPR*, pages 6982–6991. IEEE, 2020. 2

[38] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *ICCV*, pages 552–561. IEEE, 2019. 3, 4

[39] Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. Prototype completion with primitive knowledge for few-shot learning. In *CVPR*, pages 3754–3762, 2021. 2

[40] Lei Zhang, Guanyu Gao, and Huaizheng Zhang. Spatial-temporal federated learning for lifelong person re-identification on distributed edges. *arXiv:2207.11759*, 2022. 6

[41] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *CVPR*, pages 13205–13214. IEEE, 2020. 2

[42] Bo Zhao, Shixiang Tang, Dapeng Chen, Hakan Bilen, and Rui Zhao. Continual representation learning for biometric identification. In *WACV*, pages 1197–1207. IEEE, 2021. 5, 6

[43] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, pages 907–915. IEEE, 2017. 6

[44] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124. IEEE, 2015. 6

[45] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 129(4):1106–1120, 2021. 3

[46] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *CVPR*, pages 5871–5880. IEEE, 2021. 1, 2, 3, 4

[47] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *CVPR*, pages 6801–6810. IEEE, 2021. 3

[48] Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *CVPR*, pages 9296–9305. IEEE, 2022. 2

[49] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *CVPR*, pages 11444–11453. IEEE, 2023. 3