IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 December 2024, accepted 29 January 2025, date of publication 7 February 2025, date of current version 12 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3539927

RESEARCH ARTICLE

Lifelong Person Search

JAE-WON YANG[1], (Graduate Student Member, IEEE), SEUNGBIN HONG[2], AND JAE-YOUNG SIM[1,2], (Member, IEEE)

[1]Department of Electrical Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea
[2]Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea

Corresponding author: Jae-Young Sim (jysim@unist.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by Korea Government [Ministry of Science and ICT (MSIT)] under Grant RS-2024-00392536; in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by Korea Government (MSIT) (Leading Generative AI Human Resources Development) under Grant IITP-2024-RS-2024-00360227; in part by the Artificial Intelligence Gradate School Program, Ulsan National Institute of Science of Technology, under Grant RS-2020-II201336; and in part by the Artificial Intelligence Innovation Hub under Grant RS-2021-II212068.

ABSTRACT Person search is the task to localize a query person in gallery datasets of scene images. Existing methods have been mainly developed to handle a single target dataset only, however diverse datasets are continuously given in practical applications of person search. In such cases, they suffer from the catastrophic knowledge forgetting in the old datasets when trained on new datasets. In this paper, we first introduce a novel problem of lifelong person search (LPS) where the model is incrementally trained on the new datasets while preserving the knowledge learned in the old datasets. We propose an end-to-end LPS framework that facilitates the knowledge distillation to enforce the consistency learning between the old and new models by utilizing the prototype features of the foreground persons as well as the hard background proposals in the old domains. Moreover, we also devise the rehearsal-based instance matching to further improve the discrimination ability in the old domains by using the unlabeled person instances additionally. Experimental results demonstrate that the proposed method achieves significantly superior performance of both the detection and re-identification to preserve the knowledge learned in the old domains compared with the existing methods.

INDEX TERMS Person search, person re-identification, lifelong learning, continual learning.

I. INTRODUCTION

Person search is the technique to find the query person from the gallery sets of scene images where multiple persons usually appear simultaneously in each image. It has been drawing much attention due to its practical applicability to various real-world scenarios such as large-scale video understanding, surveillance, and augmented reality. Different from the person re-identification (re-ID) [1], [2], [3], [4], [5] that finds the query person from the sets of cropped person images, the person search is a more challenging task that first localizes the bounding boxes of person instances in the scene images and then matches the identities of the detected instances to the query person. The person search can be implemented in the two-step manner by using separately trained two sub-networks of the object detection and re-ID. However, the training of the two-step methods is usually inefficient requiring huge computational complexity, since the detection network extracts the bounding boxes for the person instances from the scene images which are then inputted to the re-ID network to retrieve the features again tailored to the re-ID task.

To overcome this issue, the end-to-end learning was introduced that jointly trains the person detection and re-ID networks. The end-to-end methods have been mainly developed in the supervised learning manner based on the assumption that both of the training and test data come from a same target dataset [6], [7], [8], [9], [10], [12]. However, in many practical real-world applications, multiple datasets are generated in different places and times that exhibit domain gaps from one another. In such cases, the existing supervised methods trained on a certain dataset usually fail to work on other datasets. Furthermore, re-training the network, whenever the target datasets are changed, suffers from the high computational complexity as well as the catastrophic forgetting [13] of the knowledge learned from the previously trained datasets.

The associate editor coordinating the review of this manuscript and approving it for publication was Ramakrishnan Srinivasan.

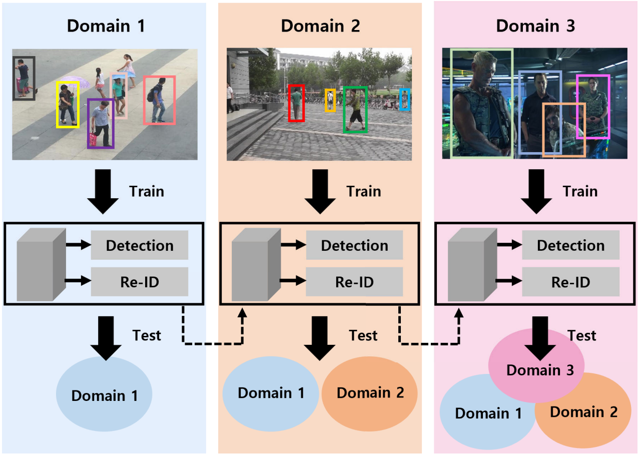

FIGURE 1. The concept of the proposed lifelong person search. New datasets on different domains are given in order. The model is incrementally trained on new domains without entire re-training on all the datasets while preserving the knowledge of old domains.

In this paper, we first introduce a new problem of lifelong person search (LPS) where the new datasets on different domains are assumed to be sequentially given in order, as shown in Fig. 1. The model is forced to be generalized to all domains while preserving the previously learned knowledge without entire re-training using all the datasets. The end-to-end LPS is more challenging compared to the lifelong object detection [14], [15] and lifelong person re-ID [16], [17], [18], [19], since it suffers from the catastrophic forgetting problem in both sub-tasks of the person detection and re-ID. Whereas the lifelong person detection is a domain-incremental task where only the same class of person is localized across different domains, the lifelong person re-ID is related to both domain-incremental and class-incremental tasks since the new person identities are additionally given from different domains. Moreover, the end-to-end person search also suffers from the task conflict problem where the person detection focuses on extracting the common representation of persons distinct from the backgrounds, but the person re-ID attempts to extract the unique representations according to the person identities. This task conflict problem becomes more serious in LPS scenario where the model is encouraged to be continuously adapted to different domains. In addition, the lifelong re-ID methods [16], [17], [18], [19] have been usually developed for full-body pedestrian images on similar domains. However, as shown in Fig. 1, the LPS considers the scene images with severely different characteristics across multiple domains, for example, diverse backgrounds, different scales and densities of persons, and even local body parts due to partial occlusion.

To address the LPS problem, we propose a novel end-to-end framework that generalizes the network to be incrementally adapted to the new domains while preserving the previously learned knowledge in the old domains. Specifically, we perform the knowledge distillation for both sub-tasks of the person detection and re-ID using the old exemplar data. We first use the prototypes, representative features of person identities, associated with the old exemplar data, to design the rehearsal-based re-ID knowledge distillation loss that enforces the consistency on the distributions of the feature similarity between the old and new models. Moreover, we also utilize the hard background proposals additionally to refine the re-ID knowledge distillation loss that alleviate the effect of inaccurately detected person proposals and extract discriminative features for re-ID more reliably. In addition, we devise the rehearsal-based instance matching loss to further improve the model's discrimination ability. We minimize the feature discrepancy between the labeled proposals in the old exemplar data and its ground truth old prototypes. We also employ the unlabeled person identities in the old domains as negative samples to preserve the knowledge effectively. Experimental results demonstrate that the proposed method preserves the knowledge of old datasets more faithfully compared with the existing methods, and therefore serves as a very promising tool for LPS.

The main contributions of this paper are summarized as follows.

- To the best of our knowledge, we first introduce a new and challenging problem of person search with lifelong learning scenario where the model is incrementally trained on the new domains while preserving the previously learned knowledge in the old domains.
- We propose an end-to-end LPS framework that jointly trains the detection and re-ID networks by using the rehearsal-based knowledge distillation loss and the instance matching loss, that alleviate the catastrophic forgetting of the old knowledge during the training on the new domains.
- We demonstrate the efficiency of the proposed method by providing comprehensive experimental results compared with the existing methods based on the lifelong learning scenario.

II. RELATED WORKS

A. PERSON SEARCH

Person search has been studied mainly in the supervised manner where the bounding boxes of person instances and the person identities are labeled in the training datasets. The two-step methods train the person detection and re-ID networks separately to prevent the conflict problem between the two tasks. Zheng et al. [20] conducted extensive experiments by training the state-of-the-art methods of the pedestrian detection and person re-ID. They also provided a benchmark PRW dataset. Chen et al. [21] proposed a segmentation masking scheme to force the re-ID network to focus on the foreground regions of the detected persons. Lan et al. [22] extracted multi-scale features to deal with the scale variation problem of person size in the scene images. Wang et al. [23] made the re-ID network more adapted to the detection results by composing the training set with the person images cropped by the pre-trained detection network and the person images cropped by using the bounding box labels. Ke et al. [24] performed a data augmentation scheme that shifts the locations of the ground truth bounding boxes for the re-ID network training.

The end-to-end methods jointly train the person detection and re-ID networks. Xiao et al. [25] firstly proposed an end-to-end person search network and provided a benchmark CUHK-SYSU dataset. Chen et al. [26] employed the background features as negative samples to train the re-ID network. Chen et al. [6] separated the feature embedding into the norm and angle which are used as a detection confidence score and an identity feature, respectively. Zhang et al. [27] pretrained an external re-ID network which is then used as a strong teacher model to supervise the re-ID network based on the knowledge distillation framework. Li and Miao [28] employed an additional Faster R-CNN header sequentially to extract the superior identity features from the high-quality person proposals. Han et al. [9] adaptively controlled the gradient backpropagation to train the sub-networks of the re-ID and part classification according to the quality of detection results. Lee et al. [10] suggested a feature standardization scheme and a localization aware memory updating scheme to alleviate the effect of class imbalance and inaccurately detected proposals, respectively. The transformer architectures [11] were also employed to improve the performance of person search [12], [29], [30], [31]. Recently, Oh et al. [32] assumed the training data of real target domains are not available, and proposed a domain generalizable person search method that uses only an unreal dataset for training.

On the other hand, the weakly-supervised person search has been introduced that uses the labeled bounding boxes only without using the identity labels for training [33], [34], [35], [36]. Moreover, domain-adaptation methods have been proposed to address the unsupervised person search problem where both of the bounding box and identity labels are not available [37], [38]. Note that the existing methods of person search have been usually developed considering a single target dataset only, and hence suffer from the catastrophic forgetting problem where new target datasets are continuously given in the lifelong learning scenario.

B. LIFELONG OBJECT DETECTION

Lifelong object detection methods are classified into the class-incremental approach and the domain-incremental approach. The class-incremental object detection considers a certain target dataset where the new object classes are incrementally added. Shmelkov et al. [14] first introduced the problem of catastrophic forgetting in the object detection. Adaptive distillation has been performed between the intermediate features and the output of the region proposal network based on the end-to-end framework [15], [39]. Shieh et al. [40] stored a subset of old data into the exemplar memory to alleviate the catastrophic forgetting in old classes. Liu et al. [41] focused on the most informative old knowledge by sampling the most reliable foreground prediction from the old model, which are then used as pseudo labels in a transformer based detection network, DETR [42]. Dong et al. [43] performed self-supervised learning with the DETR network where only a few labeled new object classes appear in the new data. On the other hand, object detection datasets are associated with different domains according to the variations of background, lighting, and camera viewpoint, even though they contain the same object classes. The domain-incremental object detection assumes incrementally added new domains with the same object class. Li et al. [44] used a transformer-based feature extractor to adaptively apply the classification head network to each newly added domain. Mirza et al. [45] stored the statistical changes across the domains used to perform the task on the corresponding specific domain.

C. LIFELONG PERSON RE-ID

Wu and Gong [16] first introduced the lifelong person re-ID problem and performed coherence learning for classification, distribution, and representation, respectively. Pu et al. [17] adaptively accumulated the knowledge of old domains via the instance-based similarity to improve the generalization ability. Ge et al. [18] developed a domain adaptation framework that reduces the gap between the old and new domains by using the augmented new data following the distributions of the old domains. Sun and Mu [46] selected diverse and important patches from images by using a differentiable patch sampler to preserve both the local and global relational knowledge. Huang et al. [19] developed a relation consistency learning to encourage the new model to return the consistent results of similarity ranking to that of the old model. Pu et al. [47] constructed meta reconciliation normalization layers that adaptively rectify domain-independent batch norm statistics. Yu et al. [48] proposed a knowledge transfer scheme via bi-directional learning that dynamically updates the old model while training the new model. Pu et al. [49] performed adaptive knowledge accumulation using graph convolution networks to maintain the domain knowledge and performed ranking consistency distillation that preserves the inter-instance ranking relationship. Xu et al. [50] used a pair-wise relation matrix to filter out the erroneous knowledge of the old model and transfer the refined knowledge to the new model. Xu et al. [51] generated the prototypes of old domains by modeling the instance features into multivariate gaussian distribution.

Note that the lifelong person re-ID methods work on cropped person images only where relatively large numbers of instances with a same identity are given in each mini-batch. Therefore, they cannot be directly applied to LPS where the detection and re-ID tasks are systematically interconnected during concurrent learning. We propose an end-to-end LPS framework to alleviate the knowledge forgetting in both tasks where a mini-batch is composed of a scene image and the number of identities is usually restricted due to memory constraint. Moreover, the proposed method can deal with many unlabeled instances that appear in scene images, whereas the person re-ID methods cannot. In addition, we first utilize the features of inaccurate background proposals provided

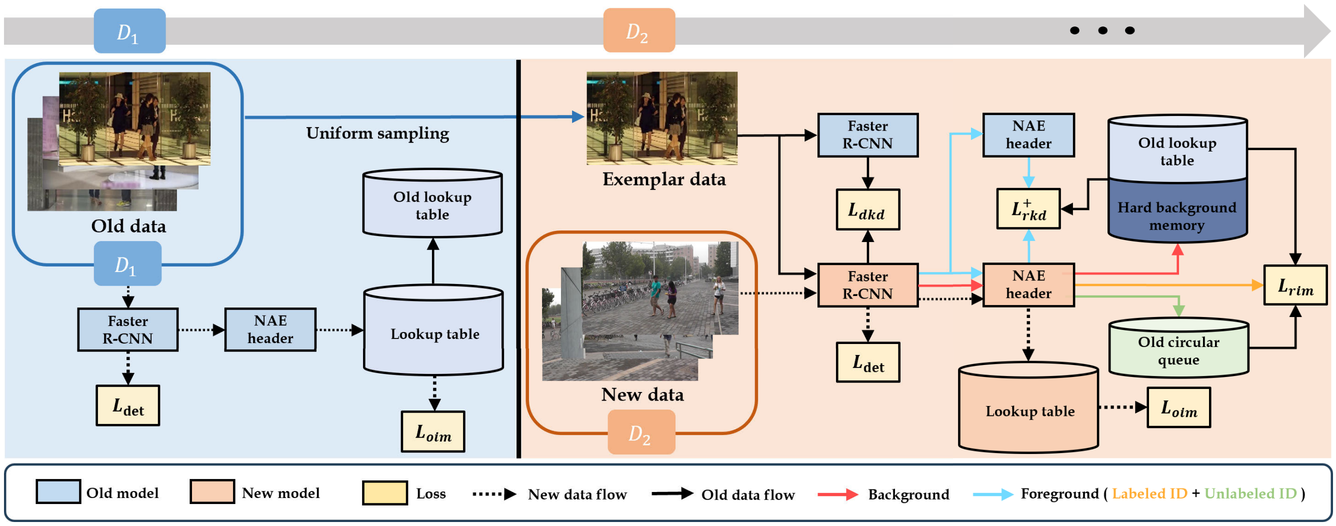

FIGURE 2. Overall framework of the proposed method.

by the detector to make the model robust to the inaccurate detection results causes by knowledge forgetting.

III. PROPOSED METHOD

Fig. 2 shows the overall architecture of the proposed end-to-end LPS framework. We use the SeqNet [28] as a baseline network which consists of the Faster R-CNN [52] and the NAE (norm-aware embedding) [6] header. Let us assume that a sequence of person search datasets in different domains are given in order, as $D_1 \rightarrow D_2 \rightarrow \ldots \rightarrow D_N$. The model is trained by using the first dataset D_1. When the new dataset D_2 is given, we regard the model trained on D_1 as the old model, and construct a new model by replicating the old model. Then the new model is trained by using D_2 and a small subset of D_1, called exemplar data, to avoid the knowledge forgetting of D_1. We also use the representative features of person identities, called prototypes [53], stored in the old look-up table (LUT) $\mathcal{T}$. Whenever a new dataset D_N is available, the old model is replaced with the new model, and the new model is re-trained by using the new data D_N as well as both the old exemplar data and the old prototypes selected from $\{D_1, \cdots, D_{N-1}\}$ to mitigate the catastrophic forgetting. We use a small subset of the old data following the typical rehearsal (replay) based methodology of lifelong learning [16], [18], [40], [53]. However, it is worth to note that we do not employ multiple old models but always have a single old model which is updated whenever a new dataset is given. The old model conveys the knowledge of the previous domains and thus the parameters of the old model are frozen during the training of the new model. We preserve the knowledge of the old data while training the model using the new data via knowledge distillation between the old and new models.

A. RE-ID KNOWLEDGE DISTILLATION

1) PROTOTYPE-BASED DISTILLATION

Existing methods of lifelong person re-ID [16], [18] perform the knowledge distillation by matching the distributions of the feature similarity between the old and new models within a mini-batch. However, it may not provide faithful results when applied to LPS, since a mini-batch is composed of relatively small numbers of scene images, and small numbers of identities accordingly, due to the memory constraints. Furthermore, multiple person instances with the same identity are rarely included in a single mini-batch according to the uniqueness prior [34].

To address this issue for LPS, we utilize the stored prototypes of person identities as informative guidance for the re-ID knowledge distillation. The prototype is computed by aggregating the features of diverse person instances with the same identity [6], [12], [25], [28]. Let $\mathcal{F}$ denotes the set of the foreground proposals in the exemplar data of a single mini-batch, that is detected by the Faster R-CNN of the new model. Let $\boldsymbol{x}_i^{\text{old}}$ and $\boldsymbol{x}_i^{\text{new}}$ be the L2-normalized features of the i-th proposal in $\mathcal{F}$, that are extracted through the NAE headers of the old and new models, respectively. We estimate the target distribution of the feature similarity of $\boldsymbol{x}_i^{\text{old}}$ compared to all the prototypes in $\mathcal{T}$, such that the probability $q_{i,k}$ associated with $\boldsymbol{x}_i^{\text{old}}$ and z_k, the k-th prototype in $\mathcal{T}$, is given by

$$q_{i,k} = \frac{\exp\left(z_k^{\mathrm{T}} \boldsymbol{x}_i^{\text{old}} / \tau_d\right)}{\sum_{z \in \mathcal{T}} \exp\left(z^{\mathrm{T}} \boldsymbol{x}_i^{\text{old}} / \tau_d\right)}. \tag{1}$$

We also estimate the predicted distribution of the feature similarity of $\boldsymbol{x}_i^{\text{new}}$ compared to all the prototypes, such that the probability $p_{i,k}$ associated with $\boldsymbol{x}_i^{\text{new}}$ and z_k is given by

$$p_{i,k} = \frac{\exp\left(z_k^{\mathrm{T}} \boldsymbol{x}_i^{\text{new}} / \tau_d\right)}{\sum_{z \in \mathcal{T}} \exp\left(z^{\mathrm{T}} \boldsymbol{x}_i^{\text{new}} / \tau_d\right)}. \tag{2}$$

Note that the predicted similarity distribution with respect to the prototypes in the old domains changes when the model is trained on the new domain, which could cause the forgetting of the re-ID knowledge learned on the old domains. Therefore, we train the new model to generate a more consistent distribution to the target distribution by employing

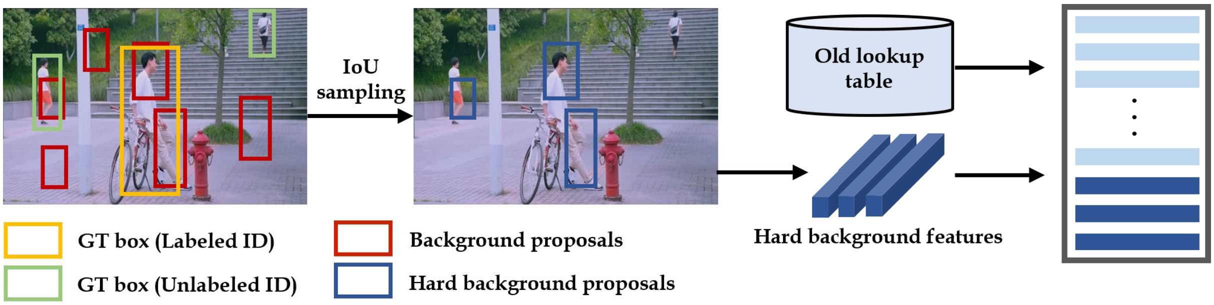

FIGURE 3. Sampling of the hard background proposals detected from the exemplar data.

a prototype-based re-ID knowledge distillation loss given by

$$\mathcal{L}_{\mathrm{rkd}} = \frac{1}{|\mathcal{F}|}\frac{1}{|\mathcal{T}|}\sum_{i\in I(\mathcal{F})}\sum_{k\in I(\mathcal{T})} q_{i,k}\log\frac{q_{i,k}}{p_{i,k}}, \tag{3}$$

where $I(\cdot)$ means the index set. By minimizing $\mathcal{L}_{\mathrm{rkd}}$, the new model is trained to yield a more consistent distribution to the target distribution with respect to the prototypes in the old domains, and eventually extracts unique and representative features of all person identities alleviating the catastrophic forgetting in re-ID.

2) HARD BACKGROUND PROPOSAL-BASED DISTILLATION

Though the prototypes in the old LUT serve as a good prior for the re-ID knowledge distillation, we further improve the performance by using the background proposals in the old domains additionally. At each iteration, the new model detects the background proposals from the scene images in the exemplar data, as depicted in the red boxes in Fig. 3. Inaccurate background proposals are often generated that partially overlap with the foreground person instances. We refer them as the hard background proposals. The hard background proposals convey the partial information of the person identities, exploited to improve the discrimination performance of the person identities.

Specifically, we sample the hard background proposals that have the higher intersection over union (IoU) scores than a certain threshold λ_b, with respect to the ground truth bounding boxes of the foreground persons, as depicted in the blue boxes in Fig. 3. Then we store the re-ID features of the hard background proposals into the feature memory $\mathcal{M}$. We re-compute the distributions of the feature similarity compared to all the prototypes in $\mathcal{T}$ as well as all the features of the hard background proposals in $\mathcal{M}$, such that the probabilities $q^{+}_{i,k}$ and $p^{+}_{i,k}$ associated with z_k, the k-th element in $\mathcal{T}\cup\mathcal{M}$, are given by

$$q^{+}_{i,k} = \frac{\exp\left(z_k^{\mathrm{T}}\boldsymbol{x}_i^{\mathrm{old}}/\tau_d\right)}{\sum_{z\in\{\mathcal{T}\cup\mathcal{M}\}}\exp\left(z^{\mathrm{T}}\boldsymbol{x}_i^{\mathrm{old}}/\tau_d\right)}, \tag{4}$$

$$p^{+}_{i,k} = \frac{\exp\left(z_k^{\mathrm{T}}\boldsymbol{x}_i^{\mathrm{new}}/\tau_d\right)}{\sum_{z\in\{\mathcal{T}\cup\mathcal{M}\}}\exp\left(z^{\mathrm{T}}\boldsymbol{x}_i^{\mathrm{new}}/\tau_d\right)}. \tag{5}$$

Accordingly, we have the refined re-ID knowledge distillation loss as

$$\mathcal{L}^{+}_{\mathrm{rkd}} = \frac{1}{|\mathcal{F}|}\frac{1}{|\mathcal{T}\cup\mathcal{M}|}\sum_{i\in I(\mathcal{F})}\sum_{k\in I(\mathcal{T}\cup\mathcal{M})} q^{+}_{i,k}\log\frac{q^{+}_{i,k}}{p^{+}_{i,k}}. \tag{6}$$

Consequently, we improve the discrimination performance of the person identities by exploiting more rich information carried by the hard background proposals. Furthermore, the features learned by additionally using the hard background proposals are more robust against the inaccurately detected person proposals, which alleviates the task conflict problem between the detection and re-ID even when the detection knowledge in the old domains is forgotten.

B. REHEARSAL-BASED INSTANCE MATCHING

The foreground and background proposals extracted from the exemplar data are used for consistent learning between the old and new models via the re-ID knowledge distillation. Note that, as depicted in the green boxes in Fig. 3, some foreground person instances have no identity labels. Such unlabeled instances can also serve as the negative samples for all the labeled identities to learn the discriminative feature representations. At the same time, the new model should be guided to minimize the feature discrepancy across the person instances in the exemplar data that have the same identity. Therefore, we also utilize the unlabeled instances in the exemplar data to further capture the re-ID knowledge in the old domains while the model is trained on the new data.

The features of the unlabeled proposals are stored in the old circular queue $\mathcal{Q}$. Let $\mathcal{F}_L$ denote the set of the labeled proposals in $\mathcal{F}$, and let $\boldsymbol{x}_i^{\mathrm{new}}$ be the feature of the i-th proposal in $\mathcal{F}_L$ extracted by the new model. We compute the probability that $\boldsymbol{x}_i^{\mathrm{new}}$ is classified into its ground truth label as

$$\rho_i = \frac{\exp\left(\boldsymbol{z}(i)^{\mathrm{T}}\boldsymbol{x}_i^{\mathrm{new}}/\tau_r\right)}{\sum_{z\in\mathcal{T}}\exp\left(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{x}_i^{\mathrm{new}}/\tau_r\right)+\sum_{y\in\mathcal{Q}}\exp\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_i^{\mathrm{new}}/\tau_r\right)}, \tag{7}$$

where $\boldsymbol{z}(i)$ means the prototype of the ground truth identity of the i-th proposal in $\mathcal{F}_L$, and $\boldsymbol{y}$ denotes the feature of the unlabeled proposals stored in $\mathcal{Q}$. We train the new model to increase the classification score of the extracted features

TABLE 1. Statistics of person search datasets.

Dataset	Training			Test		
	MovieNet-PS	CUHK-SYSU	PRW	MovieNet-PS	CUHK-SYSU	PRW
Frame	20,158	11,206	5,704	43,640	6,978	6,112
Identity	2,078	5,532	483	1,000	2,900	544
Bounding box	32,927	55,272	18,048	79,607	40,871	25,062
Query	-	-	-	1,000	2,900	2,057

by employing the rehearsal-based instance matching loss given by

$$\mathcal{L}_{\text{rim}} = -\frac{1}{|\mathcal{F}_L|} \sum_{i \in I(\mathcal{F}_L)} \log \rho_i. \tag{8}$$

By minimizing $\mathcal{L}_{\text{rim}}$, we reduce the feature discrepancy between the labeled proposal and its ground truth identity while preserving the discrimination performance in the old domains with the help of the unlabeled proposals. It is worth to note that the conventional OIM [25] loss considers the labeled identities and the unlabeled instances in a single target domain only. On the contrary, the proposed rehearsal-based loss $\mathcal{L}_{\text{rim}}$ employs the labeled identities and the unlabeled instances across the old data, aiming to preserve the discrimination performance in the old domains for lifelong learning purpose.

C. TRAINING AND INFERENCE

At the training phase, both of the person detection and re-ID networks are trained in the end-to-end manner. Note that the baseline network of SeqNet [28] also uses the losses of $\mathcal{L}_{\text{det}}$ and $\mathcal{L}_{\text{oim}}$ when training the new model by using the new dataset. To preserve the knowledge of the old domains in terms of the person detection, we additionally use the detection knowledge distillation loss $\mathcal{L}_{\text{dkd}}$ of the existing lifelong object detection method [15]. Finally, the total loss function is given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dkd}} + \mathcal{L}^{+}_{\text{rkd}} + \mathcal{L}_{\text{rim}} + \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{oim}}. \tag{9}$$

The overall training procedures of the proposed method are described in Algorithm 1.

After the new model is trained with the last dataset D_N, we discard the old model and only utilize the new model to detect the bounding boxes of the person instances and extract the re-ID features for all the datasets $\{D_1, D_2 \ldots D_N\}$ at the inference phase.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETUP

1) DATASETS

We used the three datasets to evaluate the performance of the proposed lifelong person search method. CUHK-SYSU [25] and PRW [20] are widely used for the person search task. The CUHK-SYSU dataset includes the images obtained from the street snapshots and movies, with the annotations of the bounding boxes and person identities. We set the gallery size to 100. The PRW dataset is composed of the video frames capturing a university campus by six different cameras. We also use a recently released large-scale person search dataset of MovieNet-PS [54] gathered from the 385 movie sequences. The MovieNet-PS is a challenging dataset since it includes the scene images with diverse backgrounds, illuminations, and poses of persons to reflect more realistic and challenging scenarios of person search. Moreover, there are many persons partially appearing due to the occlusion, and the person instances with the same identity often wear different clothes. The statistics of the three datasets are shown in Table 1.

Algorithm 1 Lifelong Learning Process

```
Input: Model, sequence of datasets ${D_1, D_2, \ldots, D_N}$
 Train model using dataset $D_1$
 Initialize old model with the model trained on $D_1$
 Store subset of data from $D_1$ into exemplar memory
 Store prototype($z_k$) from $D_1$ into $\mathcal{T}$
for each new dataset $D_t$ where $t = 2, \ldots, N$ do
     Replicate old model to create new model
     Freeze parameters of old model
    for epochs do
        for iterations do
            Load exemplar data and new data
            Calculate Eq.9 with new model and old model
            Update parameters of new model
        end for
    end for
     Replace old model with new model
     Store subset of data from $D_t$ into exemplar memory
     Store $z_k$ from $D_t$ into $\mathcal{T}$
end for
Output: New model
```

2) EVALUATION METRICS

We evaluated the performance of the person detection and re-ID, respectively, based on the LPS framework. We used the recall and the average precision (AP) to measure the detection performance. The recall calculates the percentage of the true positive bounding boxes where the IoU scores with respect to any ground truth bounding box are higher than 0.5. The AP computes the average precision of the bounding boxes by measuring the area under the Precision-Recall curve using the

TABLE 2. Performance of the lifelong learning evaluated on the three person search datasets with the training order of CUHK-SYSU → PRW → MovieNet-PS. The performance on each dataset is measured by using the model after the training with the last dataset is over. * means that we use the ground-truth bounding boxes for person detection. The best scores are boldfaced.

Methods	CUHK-SYSU				PRW				MovieNet-PS				Average			
	Detection		Re-ID		Detection		Re-ID		Detection		Re-ID		Detection		Re-ID	
	Recall	AP	mAP	Top-1	Recall	AP	mAP	Top-1	Recall	AP	mAP	Top-1	Recall	AP	mAP	Top-1
Joint-Train	91.3	88.1	92.9	93.7	97.8	94.8	50.9	85.4	87.1	82.9	39.9	81.2	92.1	88.6	61.2	86.8
FineTune	55.0	52.3	73.1	75.1	61.1	60.0	11.7	57.2	**91.4**	**87.1**	**40.1**	**82.3**	69.2	66.5	41.6	71.5
Det + AKA	79.5	77.9	66.0	65.1	91.9	89.6	7.4	31.2	82.3	76.1	13.0	49.6	84.6	81.2	28.8	48.6
Det + LSTKC	79.5	77.9	83.6	85.8	91.8	89.6	27.6	81.5	82.3	76.1	23.4	75.2	84.6	81.2	44.9	80.8
Det + DKP	79.5	77.9	85.8	87.3	91.8	89.6	37.8	**83.7**	82.3	76.1	21.4	73.1	84.6	81.2	48.3	81.4
Det + PTKP	79.5	77.9	82.2	83.9	91.9	89.6	33.2	72.9	82.3	76.1	20.0	66.8	84.6	81.2	45.1	74.5
Det + KRKC	79.5	77.9	83.3	85.9	91.9	89.6	29.3	80.7	82.3	76.1	25.2	75.9	84.6	81.2	45.9	80.8
Proposed	**81.4**	**79.7**	**91.2**	**92.6**	**93.9**	**91.3**	**42.2**	83.2	85.1	78.4	34.1	78.6	**86.8**	**83.1**	**55.8**	**84.8**
Det* + AKA	100	100	67.2	65.8	100	100	7.7	31.3	100	100	15.5	53.6	100	100	30.1	50.2
Det* + LSTKC	100	100	85.3	86.9	100	100	27.8	82.5	100	100	27.1	80.8	100	100	46.7	83.4
Det* + DKP	100	100	87.9	88.9	100	100	40.2	86.9	100	100	24.9	76.8	100	100	51.0	84.2
Det* + PTKP	100	100	86.1	86.4	100	100	35.8	74.1	100	100	23.4	70.3	100	100	48.4	76.9
Det* + KRKC	100	100	88.1	89.9	100	100	31.3	82.6	100	100	29.1	79.9	100	100	49.5	84.1
Proposed*	**100**	**100**	**92.8**	**93.6**	**100**	**100**	**43.4**	**84.4**	**100**	**100**	**41.3**	**84.4**	**100**	**100**	**59.2**	**87.4**

IoU scores with respect to the ground truth. We also used the mean Average Precision (mAP) and the Top-k scores for the re-ID. The mAP calculates the averaged precision of searching a query from the gallery images. It measures the area under the Precision-Recall curve using the feature similarities between the query and all the detected gallery persons. The detected bounding boxes that overlap with the ground-truth with the IoU scores higher than 0.5 are set as the true positives. The Top-k score checks whether at least one of the k-most similar candidates is a true positive or not. We adopted the Top-1 score in this work.

3) IMPLEMENTATION DETAILS

We uniformly sampled 2% of the training data from the old datasets to construct the exemplar data, as did in many literatures of lifelong person re-ID [16], [18], [48]. We store the prototypes of the persons in the exemplar data only into the old LUT, which are not updated during the training on the new domain. For the model training, we set the batch size for the exemplar data to 2 for each old domain and 5 for the new domain, respectively. We resized the input image to 1500 × 900 and applied the random horizontal flipping. We set the size of the old circular queue $\mathcal{Q}$ as 1000. We trained the model until it reaches the highest performance on the first domain, and trained the model for 5 epochs each on the other domains. Since our baseline setting achieves the best performance when the model is trained during 5 epochs for each new dataset, we also trained the model for 5 epochs on the new dataset for fair comparison to the baseline setting. The initial learning rate is set to 0.003, which is warmed up with a learning rate scheduler and further decayed by the value of 0.1 in the 3rd epoch of each new domain. For the stochastic gradient descent, we set the momentum value to 0.9 and the weight decay to 0.0005. λ_b, τ_d, and τ_r are empirically set to 0.1, 0.3 and 0.1, respectively. All our experiments were implemented using PyTorch and a single NVIDIA TITAN X GPU.

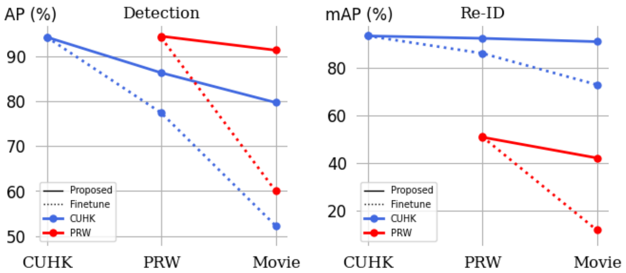

FIGURE 4. Performance in terms of the old knowledge forgetting in LPS evaluated on the old datasets of CUHK-SYSU (blue) and PRW (red), when the model is sequentially trained on each new dataset in the x-axis. The performance of the FineTune method and the proposed method are shown with the solid and dashed lines, respectively.

B. LIFELONG LEARNING PERFORMANCE

We evaluate the lifelong learning performance of the proposed method with the training order of CUHK-SYSU → PRW → MovieNet-PS and MovieNet-PS → CUHK-SYSU → PRW in Table 2 and Table 3, respectively. We first compare two different methods: Joint-Train and FineTune, implemented on our baseline network. The Joint-Train method trains the model by using all the available datasets simultaneously. The FineTune method trains the model on each dataset in order, where the model is initialized with the previously trained

TABLE 3. Performance of the lifelong learning evaluated on the three person search datasets with the training order of MovieNet-PS → CUHK-SYSU → PRW. The performance on each dataset is measured by using the model after the training with the last dataset is over. * means that we use the ground-truth bounding boxes for person detection. The best scores are boldfaced.

Methods	MovieNet-PS				CUHK-SYSU				PRW				Average			
	Detection		Re-ID		Detection		Re-ID		Detection		Re-ID		Detection		Re-ID	
	Recall	AP	mAP	Top-1	Recall	AP	mAP	Top-1	Recall	AP	mAP	Top-1	Recall	AP	mAP	Top-1
Joint-Train	87.1	82.9	39.9	81.2	91.3	88.1	92.9	93.7	97.8	94.8	50.9	85.4	92.1	88.6	61.2	86.8
FineTune	65.4	49.8	8.6	43.8	73.7	71.0	79.4	81.1	**95.7**	**93.1**	**44.8**	82.6	78.3	71.3	44.3	69.2
Det + AKA	88.2	83.3	8.2	26.4	**74.3**	**72.6**	52.3	50.8	**93.5**	**91.1**	15.4	66.1	**85.3**	**82.3**	25.3	47.8
Det + LSTKC	88.2	83.3	23.6	75.7	**74.3**	**72.6**	87.4	89.1	**93.5**	**91.1**	32.7	83.7	**85.3**	**82.3**	47.9	82.8
Det + DKP	88.2	83.3	21.7	73.8	**74.3**	**72.6**	83.4	85.2	**93.5**	**91.1**	38.0	84.9	**85.3**	**82.3**	47.7	81.3
Det + PTKP	88.2	83.3	26.5	76.3	**74.3**	**72.6**	**88.9**	**89.9**	**93.5**	**91.1**	46.3	85.7	**85.3**	**82.3**	53.9	**83.9**
Det + KRKC	88.2	83.3	17.6	64.1	**74.3**	**72.6**	86.5	88.2	**93.5**	**91.1**	**56.3**	**88.4**	**85.3**	**82.3**	53.4	80.2
Proposed	**88.9**	**84.5**	**33.8**	**77.4**	73.9	72.4	86.7	87.8	90.6	88.7	42.1	82.9	84.5	81.9	**54.2**	82.7
Det* + AKA	100	100	9.3	27.8	100	100	54.6	52.6	100	100	16.4	68.3	100	100	26.7	49.6
Det* + LSTKC	100	100	25.8	78.1	100	100	89.9	91.1	100	100	34.3	85.3	100	100	52.3	84.8
Det* + DKP	100	100	23.9	76.3	100	100	85.1	86.6	100	100	39.9	86.7	100	100	49.6	83.2
Det* + PTKP	100	100	29.0	77.5	100	100	**91.2**	**92.1**	100	100	48.8	88.4	100	100	56.3	86.0
Det* + KRKC	100	100	25.2	75.6	100	100	91.2	92.1	100	100	**59.5**	**91.9**	100	100	**58.6**	**86.5**
Proposed*	100	100	**36.7**	**80.4**	100	100	89.3	89.8	100	100	43.4	84.2	100	100	56.5	84.8

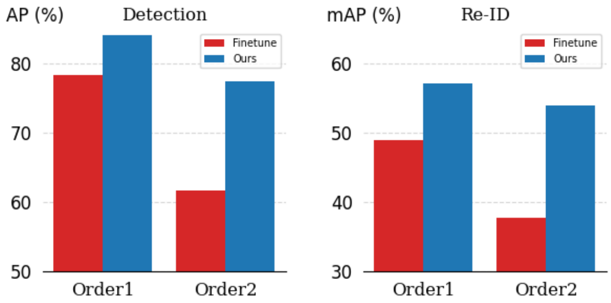

FIGURE 5. Performance of the proposed method with two different training orders. Order1: CUHK-SYSU → MovieNet-PS → PRW. Order2: PRW → CUHK-SYSU →MovieNet-PS.

weights and then re-trained by using the new dataset. At the inference phase, the performance is evaluated on every dataset by using the model trained on the last dataset.

The Joint-Train method achieves the best performance since it uses all the training datasets at once, however, it requires a huge burden of the computation as well as the storage space. We observe that the proposed method provides comparable results to the Joint-Train method and outperforms the FineTune method in terms of the averaged mAP and Top-1 score, respectively. Note that the FineTune method sequentially re-trains the model on each dataset without using the previous old datasets, and thus its performance on the last dataset, MovieNet-PS in Table 2 and PRW in Table 3, is relatively high. On the contrary, the proposed method uses a small subset of the old data to alleviate the knowledge forgetting and slightly degrades the performance on the last dataset compared to the FineTune method. However, the proposed method significantly outperforms the FineTune method in the old datasets of CUHK-SYSU and PRW in Table 2, and MovieNet-PS and CUHK-SYSU in Table 3. In addition, to verify the effect of the detection performance on the person re-ID, we measured the upper bound re-ID performance by using the ground-truth bounding boxes for the person detection during the inference phase, denoted as Det* and Proposed*. As shown in Tables 2 and 3, the averaged re-ID performance of Proposed* increases over that of the proposed method with the perfect results of detection.

Fig. 4 shows the performance evaluated on the old datasets of CUHK-SYSU (blue) and PRW (red), when the model is sequentially trained on the new datasets in the x-axis, in order. As shown by the dashed lines, both the AP and mAP scores are decreased in the FineTune method showing the knowledge forgetting effect of both the detection and re-ID tasks in the LPS scenario. However, the proposed method significantly mitigates such performance degradation as shown by the solid lines, and successfully preserves the old knowledge for LPS.

In addition, we show the performance of the proposed LPS method when the model is trained with different orders of the datasets. In Fig. 5, Order1 and Order2 represent the training orders of 'CUHK-SYSU → MovieNet-PS → PRW' and 'PRW → CUHK-SYSU → MovieNet-PS,' respectively. In both orders, we see that the proposed method still outperforms the FineTune methods in terms of the detection and re-ID performances, which indicates that the proposed method provides reliable performance regardless of the training orders.

C. COMPARISON WITH TWO-STEP METHODS

Note that we first introduce the new problem of LPS in this paper, and there is no existing method fairly comparable

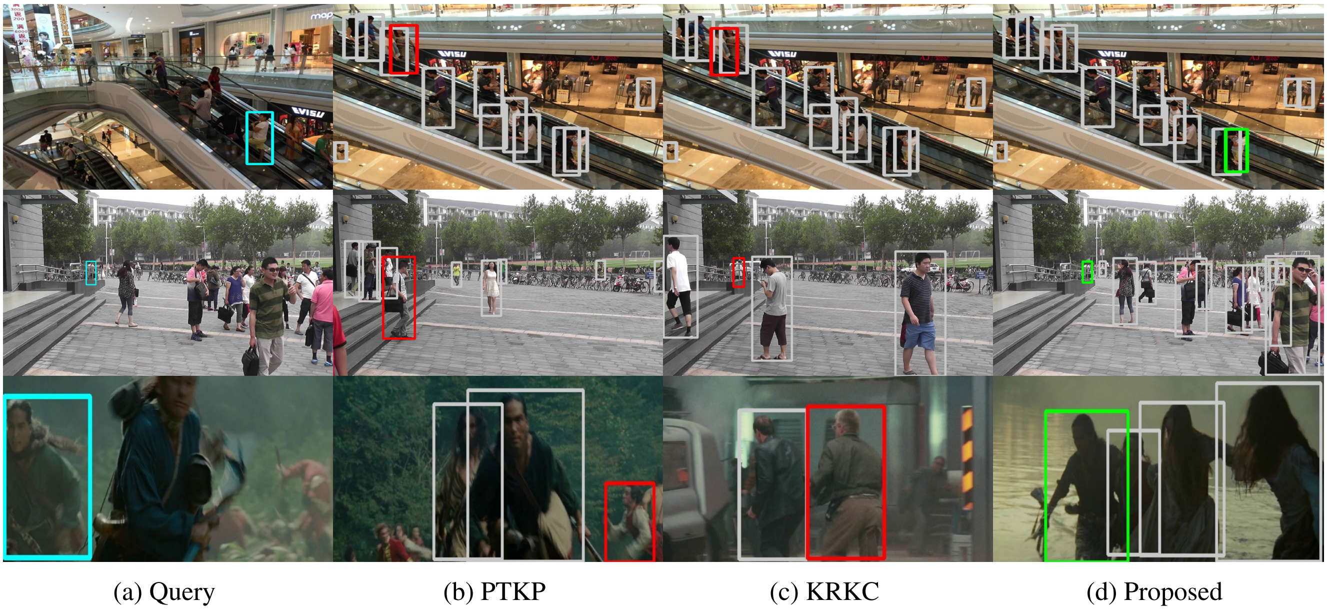
(a) Query
(b) PTKP
(c) KRKC
(d) Proposed

FIGURE 6. Qualitative comparison of the Top-1 results in the training order of CUHK-SYSU → PRW → MovieNet-PS. From top to bottom, we show the results evaluated on the test images in CUHK-SYSU, PRW, and MovieNet-PS, respectively. The query, true positive, and false positive are depicted by the blue, green, and red boxes, respectively.

to the proposed method. We attempted to conduct the additional comparative experiments by using the recent lifelong re-ID methods of AKA [17], LSTKC [50], DKP [51], PTKP [18] and KRKC [48]. The lifelong re-ID methods work on the cropped person images only and cannot be directly applied to our LPS framework that considers the scene images. We instead implemented the two-step person search framework by using the existing lifelong person re-ID methods, where the detection and re-ID networks are trained separately. We used the ResNet-50 as the backbone network for the compared methods. We also trained the detection network by using the detection knowledge distillation loss $\mathcal{L}_{\text{dkd}}$ for fair comparison.

Table 2 and Table 3 compare the quantitative performance of lifelong learning. In both Tables, we see that the proposed method achieves better performance than the two-step methods of 'Det + AKA,' 'Det + LSTKC,' 'Det + DKP,' 'Det + PTKP,' and 'Det + KRKC' in terms of the averaged re-ID performance. It means that whereas the two-step implementation of LPS by using the existing lifelong re-ID methods does not effectively reflect huge domain gaps among the person search datasets with severely different characteristics. On the other hand, the proposed end-to-end method uses prototypes as well as unlabeled instances to train the model more reliably and hence alleviate such domain gaps and being more generalizable to diverse person search datasets. In Table 3, we see that 'Det* + KRKC' shows the best performance in the last dataset of PRW, however, it suffers from the catastrophic forgetting due to a large domain gap and degrades the performance on the first dataset of MovieNet-PS.

We also compared the upper-bound performance of the person re-ID, Det* and Proposed*. Note that all the methods yield the perfect performance of detection in terms of the recall and AP, but the proposed method achieves better performance of re-ID compared with all of the two-step methods in Table 2 and most of the two-step methods in Table 3.

Fig. 6 visualizes the qualitative results of the proposed method and the two-step methods. We observe that both of PTKP [18] and KRKC [48] fail to find the query persons correctly in the challenging cases, for example, with the occluded persons and/or relatively small bounding boxes in the scene images. However, the proposed method successfully matches the query persons even in such cases, demonstrating the robustness to diverse LPS scenarios.

It is worth to note that the two-step person search framework trains the backbone network when training the detection and re-ID networks, respectively, and hence requires high computational complexity and huge memory space. In contrary, the proposed end-to-end method shares the backbone network between the jointly trained detection and re-ID networks, and is a more promising tool for LPS considering practical real-world applications.

D. ABLATION STUDY

1) LOSSES

In Table 4, we first evaluate the effect of the three losses, the detection knowledge distillation loss $\mathcal{L}_{\text{dkd}}$, the re-ID knowledge distillation loss $\mathcal{L}^{+}_{\text{rkd}}$, and the rehearsal-based instance matching loss $\mathcal{L}_{\text{rim}}$, respectively. Note that detaching all the losses is the same as the FineTune method since the proposed losses are designed to work only when the old data are available. We see that adding each loss improves the performance, respectively. Specifically, when we use $\mathcal{L}_{\text{dkd}}$, the detection performance is largely increased from that

TABLE 4. Effect of the losses used in the proposed method. The performance on each dataset is measured by using the model after the training with the last dataset is over. The best scores are boldfaced.

Methods			CUHK-SYSU				PRW				MovieNet-PS				Average			
			Detection		Re-ID		Detection		Re-ID		Detection		Re-ID		Detection		Re-ID	
$\mathcal{L}_{\text{dkd}}$	$\mathcal{L}_{\text{rkd}}^{+}$	$\mathcal{L}_{\text{rim}}$	Recall	AP	mAP	Top-1	Recall	AP	mAP	Top-1	Recall	AP	mAP	Top-1	Recall	AP	mAP	Top-1
			55.0	52.3	73.1	75.1	61.1	60.0	11.7	57.2	**91.4**	**87.1**	**40.1**	**82.3**	69.2	66.5	41.6	71.5
✓			79.5	77.9	85.1	86.8	91.9	89.6	20.7	74.9	82.3	76.1	34.1	78.7	84.6	81.2	46.6	80.1
✓	✓		81.2	79.4	90.6	91.9	92.9	90.3	39.4	82.6	85.0	78.1	34.8	79.0	86.4	82.6	55.0	84.5
✓		✓	81.0	79.2	89.4	90.9	92.8	90.4	36.5	80.6	83.7	76.7	33.5	78.7	85.8	82.1	53.1	83.4
✓	✓	✓	**81.4**	**79.7**	**91.2**	**92.6**	**93.9**	**91.3**	**42.2**	**83.2**	85.1	78.4	34.1	78.6	**86.8**	**83.1**	**55.8**	**84.8**

TABLE 5. Performance comparison when using the transformer based method of COAT [12] as the backbone network.

Methods	CUHK-SYSU		PRW		MovieNet-PS	
	mAP	Top-1	mAP	Top-1	mAP	Top-1
FineTune	68.5	71.3	11.6	54.8	**38.2**	**82.4**
Proposed	**92.4**	**91.1**	**40.7**	**83.0**	36.0	78.7

TABLE 6. Effect of using the prototype features in the old LUT for re-ID knowledge distillation.

Methods	CUHK-SYSU		PRW		MovieNet-PS	
	mAP	Top-1	mAP	Top-1	mAP	Top-1
Intra-batch	90.2	91.7	40.9	82.3	33.3	78.5
Old prototype	**91.2**	**92.6**	**42.2**	**83.2**	**34.1**	**78.6**

of the FineTune method on the old datasets, and the re-ID performance is also increased accordingly. On the other hand, $\mathcal{L}_{\text{rkd}}^{+}$ and $\mathcal{L}_{\text{rim}}$ slightly increase the detection performance of using $\mathcal{L}_{\text{dkd}}$, but significantly improve the re-ID performance on the old datasets of CUHK-SYSU and PRW by huge margins, demonstrating the effectiveness to preserve the re-ID knowledge in the old domains.

2) BASELINE NETWORK

It is worth to note that the proposed method can be applied to any baseline network of person search. We conducted the additional experiment by implementing the proposed method on the transformer based architecture of COAT [12]. Table 5 shows the results where we see that the proposed method significantly improves the performance compared to that of the FineTune method.

3) OLD PROTOTYPE-BASED KNOWLEDGE DISTILLATION

Table 6 shows the effect of using the old prototype features for re-ID knowledge distillation. The conventional method, Intra-batch, estimates the distributions of the feature similarity with respect to all the detected proposals within a mini-batch. On the other hand, the proposed method matches the distributions of the feature similarity by using the prototype features of all identities stored in the old LUT, and thus provides better results than the Intra-batch scheme.

4) EXEMPLAR DATA SAMPLING

Table 7 shows the results of using different sampling schemes to compose the exemplar data from the old datasets. 'Max BBox' samples the images that have the top 2% largest numbers of ground truth bounding boxes. 'Max ID' samples the images that have the top 2% largest numbers of person instances with identity labels. 'Random' samples 2% images randomly from the old datasets. We see that different sampling schemes provide similar performances to one another, and selected the uniform sampling that yields a slightly better performance of the old knowledge preservation compared to the other ones.

TABLE 7. Effect of the exemplar data sampling schemes.

Methods	CUHK-SYSU		PRW		MovieNet-PS	
	mAP	Top-1	mAP	Top-1	mAP	Top-1
Max BBox	**91.6**	92.3	39.0	83.0	34.0	**80.9**
Max ID	91.3	92.2	39.5	82.8	**34.8**	80.0
Random	90.6	92.0	40.8	83.2	33.8	80.2
Uniform	91.2	**92.6**	**42.2**	**83.2**	34.1	78.6

TABLE 8. Effect of the old data sampling ratio.

Methods	CUHK-SYSU		PRW		MovieNet-PS	
	mAP	Top-1	mAP	Top-1	mAP	Top-1
1%	90.2	91.7	39.7	82.1	32.8	78.8
2%	91.2	**92.6**	42.2	83.2	**34.1**	78.6
5%	91.3	92.3	44.2	83.0	33.6	**79.7**
10%	**91.4**	92.5	**45.2**	83.5	32.9	78.9
20%	91.4	92.2	43.9	**84.5**	32.2	78.5

Table 8 shows the performance variation according to the exemplar memory size by changing the old data sampling ratio, where we select 2% of the sampling ratio in this work.

E. LIMITATION

The proposed method stores 2% of the old data into the exemplar memory the typical rehearsal (replay) based methodology of lifelong learning [16], [18], [40], [53]. Therefore, the size of the exemplar memory increases as we have more and more datasets. As a future research topic, we will investigate other methodologies of lifelong learning that address the limitation of using the exemplar memory. In addition, the generalized knowledge obtained by multi-modal learning such as [55] and [56] may improve the performance of person search.

V. CONCLUSION

In this paper, we proposed a novel LPS framework where the model needs to be incrementally trained on the new datasets while preserving the knowledge of the old datasets. We implemented the knowledge distillation between the old and new models based on the rehearsal methodology by using the representative prototype features of the labeled foreground persons as well as the hard background proposals in the old exemplar data. We also designed the rehearsal-based instance matching loss to improve the discrimination ability by using the unlabeled person instances in addition to the prototype features. Experimental results evaluated on three datasets of person search showed that the proposed method achieves significantly better performance of lifelong learning compared with the existing methods, and successfully prevents the knowledge forgetting in the old domains. We expect this pioneering work would encourage further research for practical LPS applications, such as autonomous vehicles and robot navigation where new environments with various people are continuously encountered.

REFERENCES

[1] Y. Gong, L. Wang, Y. Li, and A. Du, ''A discriminative person re-identification model with global-local attention and adaptive weighted rank list loss,'' *IEEE Access*, vol. 8, pp. 203700–203711, 2020.

[2] G. Tang, X. Gao, Z. Chen, and H. Zhong, ''Graph neural network based attribute auxiliary structured grouping for person re-identification,'' *IEEE Access*, early access, Mar. 30, 2021, doi: 10.1109/ACCESS.2021.3069915.

[3] C. Zhu, W. Zhou, Y. Zhu, and J. Ma, ''Neighboring-part dependency mining and feature fusion network for person re-identification,'' *IEEE Access*, vol. 11, pp. 49760–49771, 2023.

[4] 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. (CVPR)*, Jun. 2015, pp. 3908–3916.

[5] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, ''A strong baseline and batch normalization neck for deep person re-identification,'' *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020.

[6] D. Chen, S. Zhang, J. Yang, and B. Schiele, ''Norm-aware embedding for efficient person search,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12612–12621.

[7] H. Kim, S. Joung, I.-J. Kim, and K. Sohn, ''Prototype-guided saliency feature learning for person search,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4865–4874.

[8] Y. Yan, J. Li, J. Qin, S. Bai, S. Liao, L. Liu, F. Zhu, and L. Shao, ''Anchor-free person search,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7690–7699.

[9] B.-J. Han, K. Ko, and J.-Y. Sim, ''End-to-end trainable trident person search network using adaptive gradient propagation,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 925–933.

[10] S. Lee, Y. Oh, D. Baek, J. Lee, and B. Ham, ''OIMNet++: Prototypical normalization and localization-aware learning for person search,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2022, pp. 621–637.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ''An image is worth 16×16 words: Transformers for image recognition at scale,'' in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2021.

[12] R. Yu, D. Du, R. LaLonde, D. Davila, C. Funk, A. Hoogs, and B. Clipp, ''Cascade transformers for end-to-end person search,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7267–7276.

[13] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwińska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, ''Overcoming catastrophic forgetting in neural networks,'' *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.

[14] K. Shmelkov, C. Schmid, and K. Alahari, ''Incremental learning of object detectors without catastrophic forgetting,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3400–3409.

[15] C. Peng, K. Zhao, and B. C. Lovell, ''Faster ILOD: Incremental learning for object detectors based on faster RCNN,'' *Pattern Recognit. Lett.*, vol. 140, pp. 109–115, Dec. 2020.

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

[17] N. Pu, W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew, ''Lifelong person re-identification via adaptive knowledge accumulation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7901–7910.

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

[19] Z. Huang, Z. Zhang, C. Lan, W. Zeng, P. Chu, Q. You, J. Wang, Z. Liu, and Z.-J. Zha, ''Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14288–14297.

[20] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, ''Person re-identification in the wild,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1367–1376.

[21] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, ''Person search via a mask-guided two-stream CNN model,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 734–750.

[22] X. Lan, X. Zhu, and S. Gong, ''Person search by multi-scale matching,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 536–552.

[23] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, ''TCTS: A task-consistent two-stage framework for person search,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11952–11961.

[24] X. Ke, H. Liu, W. Guo, B. Chen, Y. Cai, and W. Chen, ''Joint sample enhancement and instance-sensitive feature learning for efficient person search,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7924–7937, Nov. 2022.

[25] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, ''Joint detection and identification feature learning for person search,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3415–3424.

[26] D. Chen, S. Zhang, W. Ouyang, J. Yang, and B. Schiele, ''Hierarchical online instance matching for person search,'' in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 10518–10525.

[27] X. Zhang, X. Wang, J.-W. Bian, C. Shen, and M. You, ''Diverse knowledge distillation for end-to-end person search,'' in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 4, pp. 3412–3420.

[28] Z. Li and D. Miao, ''Sequential end-to-end network for efficient person search,'' in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 3, pp. 2011–2019.

[29] J. Cao, Y. Pang, R. M. Anwer, H. Cholakkal, J. Xie, M. Shah, and F. S. Khan, ''PSTR: End-to-end one-step person search with transformers,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9458–9467.

[30] M. Fiaz, H. Cholakkal, R. M. Anwer, and F. Shahbaz Khan, ''SAT: Scale-augmented transformer for person search,'' in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4820–4829.

[31] X. Yang, M. Tian, N. Wang, and X. Gao, ''Unleashing the feature hierarchy potential: An efficient tri-hybrid person search model,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 11, pp. 11551–11563, Nov. 2024.

[32] M. Oh, D. Kim, and J.-Y. Sim, ''Domain generalizable person search using unreal dataset,'' in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 5, pp. 4361–4368.

[33] C. Han, K. Su, D. Yu, Z. Yuan, C. Gao, N. Sang, Y. Yang, and C. Wang, ''Weakly supervised person search with region Siamese networks,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12006–12015.

[34] B.-J. Han, K. Ko, and J.-Y. Sim, ''Context-aware unsupervised clustering for person search,'' in *Proc. Brit. Mach. Vis. Conf.*, Jan. 2021.

[35] Y. Yan, J. Li, S. Liao, J. Qin, B. Ni, K. Lü, and X. Yang, ''Exploring visual context for weakly supervised person search,'' in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 3, pp. 3027–3035.

[36] B. Wang, Y. Yang, J. Wu, G.-J. Qi, and Z. Lei, ''Self-similarity driven scale-invariant learning for weakly supervised person search,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1813–1822.

[37] J. Li, Y. Yan, G. Wang, F. Yu, Q. Jia, and S. Ding, ''Domain adaptive person search,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ECCV)*, Jan. 2022, pp. 302–318.

[38] M. K. Almansoori, M. Fiaz, and H. Cholakkal, "DDAM-PS: Diligent domain adaptive mixer for person search," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 6688–6697.

[39] Y. Hao, Y. Fu, Y.-G. Jiang, and Q. Tian, "An end-to-end architecture for class-incremental object detection with knowledge distillation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1–6.

[40] J.-L. Shieh, Q. M. U. Haq, M. A. Haq, S. Karam, P. Chondro, D.-Q. Gao, and S.-J. Ruan, "Continual learning strategy in one-stage object detection framework based on experience replay for autonomous driving vehicle," *Sensors*, vol. 20, no. 23, p. 6777, Nov. 2020.

[41] Y. Liu, B. Schiele, A. Vedaldi, and C. Rupprecht, "Continual detection transformer for incremental object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23799–23808.

[42] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2020, pp. 213–229.

[43] N. Dong, Y. Zhang, M. Ding, and G. H. Lee, "Incremental-DETR: Incremental few-shot object detection via self-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 1, pp. 543–551.

[44] D. Li, G. Cao, Y. Xu, Z. Cheng, and Y. Niu, "Technical report for ICCV 2021 challenge SSLAD-Track3B: Transformers are better continual learners," 2022, *arXiv:2201.04924*.

[45] M. J. Mirza, M. Masana, H. Possegger, and H. Bischof, "An efficient domain-incremental learning approach to drive in all weather conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3001–3011.

[46] Z. Sun and Y. Mu, "Patch-based knowledge distillation for lifelong person re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 696–707.

[47] N. Pu, Y. Liu, W. Chen, E. M. Bakker, and M. S. Lew, "Meta reconciliation normalization for lifelong person re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 541–549.

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

[49] 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.

[50] K. Xu, X. Zou, and J. Zhou, "LSTKC: Long short-term knowledge consolidation for lifelong person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 14, pp. 16202–16210.

[51] K. Xu, X. Zou, Y. Peng, and J. Zhou, "Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16604–16613.

[52] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015.

[53] 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. (CVPR)*, Jul. 2017, pp. 2001–2010.

[54] J. Qin, P. Zheng, Y. Yan, R. Quan, X. Cheng, and B. Ni, "Movienet-PS: A large-scale person search dataset in the wild," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[55] W. Zhang, P. Janson, R. Aljundi, and M. Elhoseiny, "Overcoming generic knowledge loss with selective parameter update," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 24046–24056.

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

JAE-WON YANG (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Kookmin University, Seoul, South Korea, in 2021. He is currently pursuing the Ph.D. degree in electrical engineering with Ulsan National Institute of Science and Technology, Ulsan, South Korea. His research interests include computer vision and deep learning.

SEUNGBIN HONG received the B.S. degree in computer science and engineering from Ulsan National Institute of Science of Technology (UNIST), South Korea, in 2023, where she is currently pursuing the Ph.D. degree in artificial intelligence. Her research interests include computer vision and deep learning.

JAE-YOUNG SIM (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 1999, 2001, and 2005, respectively. From 2005 to 2009, he was a Research Staff Member with the Samsung Advanced Institute of Technology, Samsung Electronics Company Ltd. In 2009, he joined the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea, where he is currently a Professor with the Graduate School of Artificial Intelligence and the Department of Electrical Engineering. From 2020 to 2021, he was a Visiting Researcher with the University of California at San Diego, USA. From 2021 to 2024, he was the Dean of the College of Information and Biotechnology, UNIST. He is also the Head of the Graduate School of Artificial Intelligence, UNIST. He published over 80 papers of international journals and conferences. His research interests include image processing, computer vision, and machine learning. He is an Associate Editor of *Journal of Visual Communication and Image Representation*.