

# Prompting Continual Person Search

Pengcheng Zhang  
School of Computer Science and  
Engineering, State Key Laboratory of  
Complex & Critical Software  
Environment, Jiangxi Research  
Institute, Beihang University  
Beijing, China  
pengchengz@buaa.edu.cn

Xiaohan Yu  
School of Computing, Macquarie  
University  
Sydney, Australia  
xiaohan.yu@mq.edu.au

Xiao Bai\*  
School of Computer Science and  
Engineering, State Key Laboratory of  
Complex & Critical Software  
Environment, Jiangxi Research  
Institute, Beihang University  
Beijing, China  
baixiao@buaa.edu.cn

Jin Zheng\*  
School of Computer Science and  
Engineering, State Key Laboratory of  
Complex & Critical Software  
Environment, Jiangxi Research  
Institute, Beihang University  
Beijing, China  
jinzheng@buaa.edu.cn

Xin Ning  
Institute of Semiconductors, Chinese  
Academy of Sciences  
Beijing, China  
ningxin@semi.ac.cn

## ABSTRACT

The development of person search techniques has been greatly promoted in recent years for its superior practicality and challenging goals. Despite their significant progress, existing person search models still lack the ability to continually learn from increasing real-world data and adaptively process input from different domains. To this end, this work introduces the continual person search task that sequentially learns on multiple domains and then performs person search on all seen domains. This requires balancing the stability and plasticity of the model to continually learn new knowledge without catastrophic forgetting. For this, we propose a Prompt-based Continual Person Search (PoPS) model in this paper. First, we design a compositional person search transformer to construct an effective pre-trained transformer without exhaustive pre-training from scratch on large-scale person search data. This serves as the fundamental for prompt-based continual learning. On top of that, we design a domain incremental prompt pool with a diverse attribute matching module. For each domain, we independently learn a set of prompts to encode the domain-oriented knowledge. Meanwhile, we jointly learn a group of diverse attribute projections and prototype embeddings to capture discriminative domain attributes. By matching an input image with the learned attributes across domains, the learned prompts can be properly selected for model inference. Extensive experiments are conducted

to validate the proposed method for continual person search. The source code is available at <https://github.com/PatrickZad/PoPS>.

## CCS CONCEPTS

• **Computing methodologies** → **Object identification.**

## KEYWORDS

Visual Prompt, Continual Learning, Person Search

### ACM Reference Format:

Pengcheng Zhang, Xiaohan Yu, Xiao Bai, Jin Zheng, and Xin Ning. 2024. Prompting Continual Person Search. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24)*, October 28–November 1, 2024, Melbourne, Australia. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3664647.3681240>

## 1 INTRODUCTION

Person search [58, 59, 69] aims to localize a target person in a gallery of uncropped scene images. It has attracted increasing research interest for its practicability and challenging goals. Existing works for person search have focused on boosting the performance under typical fully [1, 6, 30, 66] or weakly [16, 50, 60] supervised scenarios, and exploring domain adaptation [27] or generalization [35] methods. Despite their significant progress, these works learn only on a fixed and limited set of data while the real-world data is continually accumulating from different domains. To this end, we propose to explore the continual person search (CPS) problem that learns from sequentially incoming domains and adaptively completes the person search task for any learned domain (see Figure 1).

A major challenge for enabling CPS is to balance the stability and plasticity of the model to consistently adapt to new domains without catastrophic forgetting of seen domains. Recent works [14, 44, 55, 56, 56, 57] for continual image classification have drawn inspiration from visual prompt tuning [22] to employ a frozen pre-trained transformer [11, 49] to guarantee model stability and

\*Corresponding author

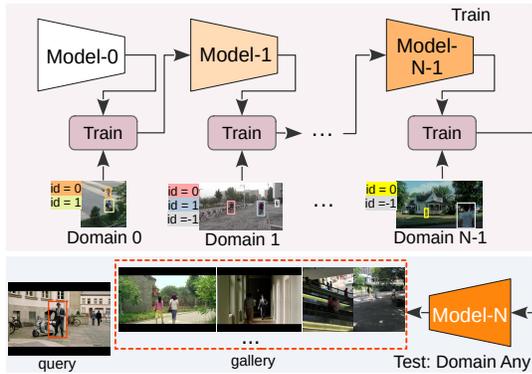
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

ACM MM, 2024, Melbourne, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0686-8/24/10

<https://doi.org/10.1145/3664647.3681240>



**Figure 1: Illustration of the continual person feature problem.**

expandable visual prompts to encode domain-oriented knowledge for plasticity. In this way, the inference of the models relies on properly selecting learned prompts to classify an image from any seen domain. As the transformers are pre-trained to learn image-level object visual representations, it is natural to incorporate those models to tackle the classification tasks. However, the models are not compatible with person search as the task requires jointly localizing and extracting instance-level features of persons in the scene image. A straightforward solution for this is to collect large-scale scene images of persons and pre-train a re-designed person search transformer from scratch. Yet this can be expensive due to (1) collecting and annotating sufficient data, *e.g.* 14M images as in ImageNet-21K [9], and (2) performing large-scale pre-training which may require a dozen large-memory GPUs running for several days.

Besides, previous prompt-based continual learning methods mainly tackle the class incremental learning [14, 23, 44, 56, 57] and domain incremental learning [23, 55, 57] scenarios. Given that the former learns from sequential datasets with disjoint semantic space and the latter assumes all learning datasets share the same semantic space [48, 54], CPS is more closely related to the domain incremental learning scenario. However, the learning domains [36] in those works are with clear boundaries (*e.g.* sketch image domain vs realistic image domain) which ease the adaptive selection of learned prompts during inference. In contrast, the domain gap between person search datasets can be ambiguous (*e.g.* both CUHK-SYSU [58] and PRW [69] contain real-world images). This further raises a challenge to robustly capture the domain-specific attributes of an input image for properly selecting learned prompts to complete the person search task.

To tackle the aforementioned problems, we propose a **Prompt-based Continual Person Search (PoPS)** model in this work. Specifically, we design a compositional person search transformer that employs an existing hierarchical vision transformer, *e.g.* Swin [33], and expand the transformer with a Simple Feature Pyramid [29] to enable person localization. The vision transformer is pre-trained on the ImageNet-22K [9] dataset and is publicly available. We then only train the Simple Feature Pyramid on a moderate number of person detection data to form a pre-trained person search transformer. This makes better use of existing pre-trained transformers

to reduce the consumption of energy and resources of large-scale pre-training from scratch. It also requires less data to optimize such a lightweight detection sub-network.

On top of the proposed person search transformer, we design a domain incremental prompt pool with diverse attribute matching to enable CPS. The visual prompts are learned independently for each domain in the continual learning procedure similar to S-Prompts [55]. To capture domain-specific attributes for selecting learned prompts, we jointly learn a group of attribute projections and prototype embeddings. Similar to [44, 55–57], a pre-trained transformer is employed to extract the global feature of an input image as the query embedding. We then use the attribute projection embedding to uncover the domain attribute in the query embedding and learn to match the attribute with the correlated attribute prototype embedding. By enforcing the attribute projection and prototype embeddings to be diverse, this diverse attribute matching mechanism is capable of capturing discriminative domain attributes. Therefore, the correlated prompts can be selected by measuring the similarity between a query embedding and learned attributes across different domains.

To summarize, this paper makes the following contributions:

- We for the first time propose the continual person search problem. A Prompt-based Continual Person Search model is presented to consistently learn to adapt new person search tasks without catastrophic forgetting.
- By constructing a compositional person search transformer, we reduce the cost of pre-training for prompt-based continual person search from scratch. A domain incremental prompt pool with diverse attribute matching is proposed to adaptively reuse learned prompts by measuring the similarity between input images and learned attributes across domains.
- Extensive experiments are conducted to understand the effectiveness of the proposed modules for continual person search.

## 2 RELATED WORK

**Person Search.** The standard supervised learning of person search has been widely explored to achieve effective person search. Zheng et al. [69] first explores combining popular person detector and person re-identification (ReID) models for person search, resulting in a two-step mechanism that first detects and crops person images and then retrieves a target across the cropped images. Following this mechanism, recent works obtained improved performance by enhanced person Re-ID features [5, 26] or designing target-conditioned person detectors [10, 52]. To improve the efficiency of the two-step paradigm, Xiao et al. [58] proposed an end-to-end method to perform person search by a unified model. Following works [4, 6, 18, 47] further explored to effectively balance the multiple training objectives as one-step person search is a multi-task learning problem. Other methods [32, 34, 63] employed the context prior knowledge to match different persons across images. It is also practical to improve model efficiency by introducing lightweight detector architectures [62, 68], or boost the model performance by designing a stronger detection sub-network [30]. Inspired by recent advances in vision transformers [11, 49, 70], recent works [1, 66]

obtained more discriminative person features with well-designed person search transformers. Recent works also explored the weakly supervised person search problem [17, 51, 61] to train a person search model with only person bounding box annotations, and unsupervised domain adaptive person search [28] that pre-trains on a labeled source domain and then adapts to an unlabeled target domain.

**Prompt-based Continual Learning.** To enable continual learning without a rehearsal buffer [3, 19], Wang et al. [57] proposed the first prompt-based continual learning method that designs a prompt pool with paired keys and prompts. The prompts help a pre-trained transformer to adapt to new tasks by visual prompt tuning [22], and the keys are learned to adaptively pick prompts for an input image. Based on this, DualPrompt [56] jointly learned task-specific expert prompts and task-shared general prompts for prompt tuning. CODA-P [44] introduced a set of prompt components and implicitly learned the attention weight for fusing the prompt components, allowing adaptive weighted prompt summation instead of selection. LGCL [25] further introduced language guidance to learn a unified semantic embedding space for continual classification. DAP [24] instead learned a prompt generation module to construct a pool-free approach. Other works [46, 53] analyzed the effect of pre-trained transformers and boosted the performance when using self-supervised pre-trained [2, 8] or moderate pre-trained transformers. Different from these works that mainly tackle class incremental learning [48, 50], S-Prompts [55] designed a simple yet effective method for domain incremental learning [48, 54].

**Lifelong ReID.** Another closely related topic to CPS is lifelong person re-identification (LReID) that learns from sequential ReID domains. For LReID, AKA [37] and MEGE [39] designed effective knowledge graphs to adaptively accumulate and reuse learned knowledge. Sun and Mu [45] proposed to adaptively choose patches for knowledge distillation. MRN [38] took a deep step into the proper batch normalization layers for LReID. Compared with these works, this work shares a similar spirit to encode and adaptively reuse learned knowledge by well-designed modules for multi-domain continual learning of person retrieval. This also eliminates the need for data-replay in many other LReID works [15, 65]. Meanwhile, this work also differs from those works as we simultaneously tackle the continual learning of person detection and present a unified framework to enable the continual learning of the two subtasks to collaboratively complete the CPS task.

## 3 METHOD

### 3.1 Problem Definition and Overview

In CPS, a unified model is trained sequentially on  $T$  domains of person search and tested on an arbitrary domain without knowing the domain identity. Denoted by  $\mathcal{D} = \{(D_{tr}^i, D_{te}^i)\}_{i=1}^T$  the learnable domains where  $D_{tr}^i$  and  $D_{te}^i$  are the train and test sets of domain  $i$ , respectively.  $D_{tr}^i = \{\mathcal{X}_{tr}^i, \mathcal{Y}_{tr}^i\}$  where  $\mathcal{X}_{tr}^i$  indicates the training scene images and  $\mathcal{Y}_{tr}^i$  contains identities and bounding boxes of persons in the images. Similarly, the test set is denoted by  $D_{te}^i = \{\mathcal{X}_{te}^i, \mathcal{Y}_{te}^i\}$ .  $\mathcal{Y}_{tr}^i \cap \mathcal{Y}_{te}^i = \emptyset$  and  $D^i \cap D^{\neq i} = \emptyset$ . During training, a unified model sequentially learns on  $T$  train sets without

accessing previously seen domains. During evaluation, the model is independently tested on the  $T$  domains.

The overall architecture of the proposed PoPS model is depicted in Figure 2. We at first design a compositional person search transformer by expanding a hierarchical vision transformer [33] pre-trained on the widely-used ImageNet [9] data with a Simple Feature Pyramid [29] (see Figure 2a). As the vision transformer is naturally capable of extracting person visual features, this enables the overall model to localize appeared persons in scene images and forms an effective person search network. As the CPS problem is more closely related to the domain incremental learning task [48, 54], we propose to solve the problem by learning domain incremental visual prompts. Specifically, on top of the proposed person search network, we design a domain incremental prompt pool [44, 55–57] that independently learns visual prompts correlated with seen domains (see Figure 2b). By introducing diverse learnable attribute projection and prototype embeddings, the prompt pool learns to capture diverse domain attributes to select proper prompts at the test time (see Figure 2c). On top of the designed modules, the overall continual learning procedure can be completed in two stages: (1) pre-training on a person detection task and (2) continual learning on incoming person search tasks.

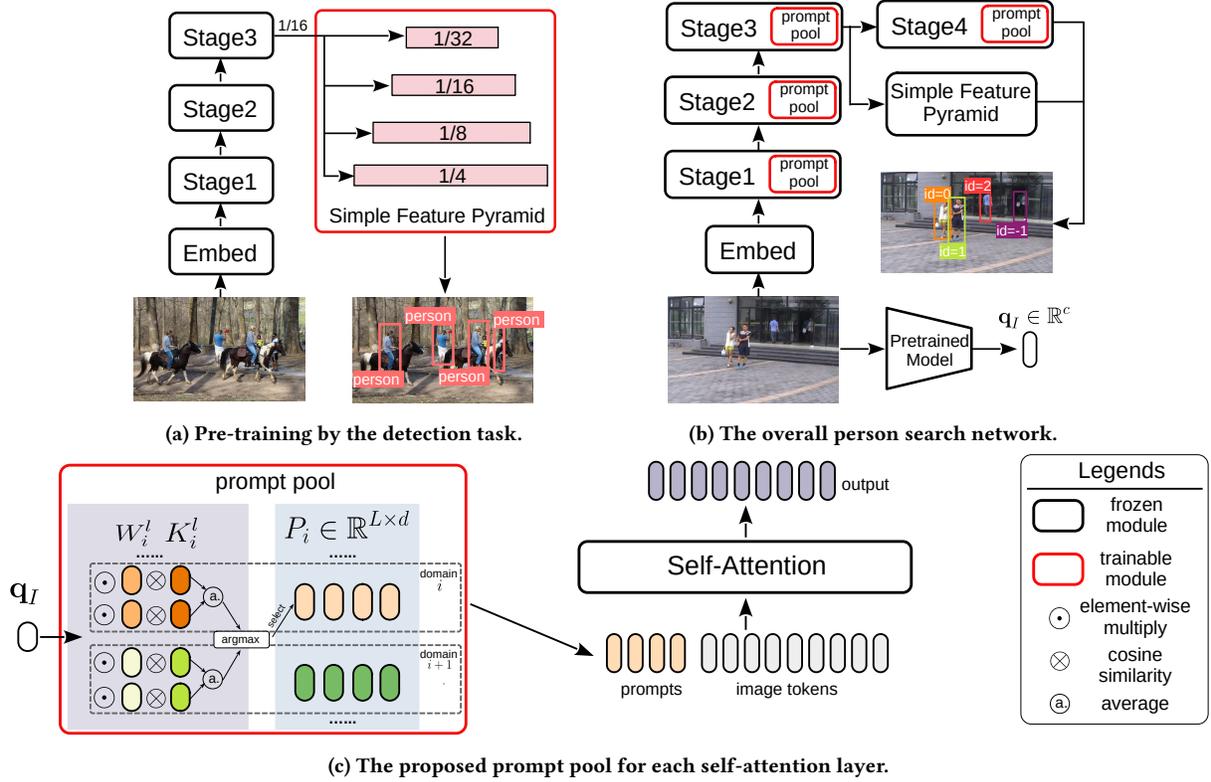
In this section, we first describe how the compositional person search transformer process input scene images to complete the task in subsection 3.2, and then introduce the person detection pre-training and person search continual learning stages in subsections 3.3 and 3.4, respectively.

### 3.2 Compositional Person Search Transformer

Largely pre-trained vision transformers [11, 49] are vital for prompt-based continual learning [44, 55–57]. Yet the models are not directly applicable to the person search task for the lack of person localization modules. It is also time- and resource-consuming to reformulate and pre-train a person search transformer from scratch. To this end, we employ a typical hierarchical vision transformer, *i.e.* Swin [33], pre-trained on the ImageNet [9] data and expand the transformer with a Simple Feature Pyramid [29] to enable person localization.

As in Figure 2a, an input image  $\mathbf{I} \in \mathbb{R}^{H_0 \times W_0 \times 3}$  is partitioned into multiple equally-sized patches and each patch is projected into a high-dimensional vector by the ‘Patch Embed’ layer, resulting in an intermediate image feature map  $\mathbf{F}_{img} \in \mathbb{R}^{H \times W \times C}$ . Afterward, Swin introduces 4 stages where each stage performs downsampling and consecutive window-based multi-head self-attention [49] on their inputs to produce deep visual representations. On top of that, we introduce the Simple Feature Pyramid [29] to enable the overall model for person localization. Following [29], we build the Simple Feature Pyramid on the 16 times downsampled feature map  $\mathbf{F}'_{img} \in \mathbb{R}^{H' \times W' \times C'}$ , *i.e.* the output of ‘Stage3’. By applying convolutions of strides  $\{2, 1, \frac{1}{2}, \frac{1}{4}\}$  on  $\mathbf{F}'_{img}$  in parallel, where the fractional strides indicate deconvolutions, we obtain image feature maps of scales  $\{\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}\}$ . Based on the multi-scale features maps, we follow Mask R-CNN [20] to predict person bounding boxes and detection confidences.

Different from the image classification [12, 33] or person re-identification tasks [64, 67] that directly extract object features from object-centered images, the person search task requires to



**Figure 2: Overview of the proposed method. (a) We first introduce a simple feature pyramid and perform detection pretraining to construct a pre-trained person search transformer. (b) We then inject the prompt pool for each transformer layer to enable CPS. The pre-trained model jointly extracts the global image feature  $q_I$  which encodes domain-related information. (c) For each self-attention layer, the prompt pool independently learns the attribute projections  $W_i^l$ , attribute prototypes  $K_i^l$ , and prompts  $P_i$  for each domain.  $q_I$  is used to match with learned domain attributes and select the prompts for test.**

extract features of instances in the scene images. To this end, we dissect Swin into two parts, *i.e.* ‘Patch Embed’ to ‘Stage3’ blocks that process the integral image feature maps, and the ‘Stage4’ block that refines instance-wise person feature maps. According to the person bounding boxes predicted by the detection sub-network, we use ROIAlign [20] to extract interpolated feature maps from the output of ‘Stage3’ within the boxes as person feature maps. We then employ the ‘Stage4’ block on the extracted feature maps to obtain refined person feature maps  $f_{roi} \in \mathbb{R}^{h \times w \times d}$ . Afterward, we follow the practice in Swin to conduct global average pooling on  $f_{roi}$  and use the pre-trained Layer Normalization layer to produce final 1D person features  $v \in \mathbb{R}^d$ .

### 3.3 Pre-training by Person Detection

Although the compositional person search transformer makes a pre-trained transformer applicable for person search, the newly added modules are randomly initialized and thus contain no prior knowledge for subsequent continual learning. To tackle this problem, we propose to pre-train only the detection sub-network to construct a pre-trained person search transformer. As the pre-trained Swin [33] is frozen in this stage, the pre-trained visual feature space is left unchanged for person feature extraction and the detection

sub-network only needs to predict person locations from the pre-trained image features. It is also worth noting that the number of the learnable parameters in this mechanism is relatively small, thus the pre-training can be completed with less data.

Specifically, we combine the training set of CrowdHuman [42] and images containing humans in the training set of MSCOCO [31] to form a person detection dataset similar to Shuai et al. [43]. In total, this collection presents 79,115 scene images that contain nearly 0.6M person instances for training the model. Compared with the pre-training (using ImageNet-21K [9] that contains 14M images) of the widely used vision transformers in prompt-based continual learning methods [44, 55–57], the proposed pre-training by person detection is more data-efficient.

Based on the collected data, we pre-train the model by conducting person detection to form a pre-trained person search transformer. As is mentioned in Subsection 3.2, we employ the ‘Patch Embed’ to ‘Stage3’ blocks of the pre-trained Swin [33] to extract image feature maps and send the feature maps to the detection sub-network to detect appeared persons. The overall training objective is formulated as

$$\mathcal{L}_{det} = \mathcal{L}_{reg} + \mathcal{L}_{cls} + \mathcal{L}_{reg}^{rpn} + \mathcal{L}_{cls}^{rpn} \quad (1)$$

where  $\mathcal{L}_{reg}$  and  $\mathcal{L}_{cls}$  are the bounding box regression and classification losses of the detection head [20],  $\mathcal{L}_{reg}^{rpn}$  and  $\mathcal{L}_{cls}^{rpn}$  are those of the Region Proposal Network [41].

We also note that the data used for this pre-training stage shares similar spirits with that in [43]. However, Shuai et al. [43] employs the data to pre-train a convolutional person search network by a self-supervised person similarity learning framework. The pre-trained model can be fine-tuned for a specific downstream task to achieve improved person search performance. Different from Shuai et al. [43], we focus on efficiently pre-training a person search transformer by a person detection task with the collected data. The pre-trained model is used to enable prompt-based continual learning without fine-tuning.

### 3.4 Continual Learning for Person Search

A key target in continual learning is to balance the stability and plasticity of the model for accumulating knowledge of new tasks without forgetting that of learned tasks. For this, recent advanced continual learning methods [23, 44, 55–57] have explored a prompt-based mechanism that exploits a largely pre-trained vision transformer with an incrementally learned prompt [22] pool. In this way, the pre-trained transformer is fixed and a set of learnable prompts are injected into the input sequence to the self-attention layer [11, 49] to adapt to a new task. Thus the continual learning problem can be solved by continually learning new visual prompts for new tasks during training and adaptively reusing proper prompts during inference. Inspired by this mechanism, we design a domain incremental prompt pool with diverse attribute matching to enable CPS.

**Domain incremental prompt pool.** On top of the pre-trained compositional person search transformer, we independently learn domain-oriented visual prompts during continual learning similar to S-Prompts [55]. Specifically for a domain  $i$ , we set a sequence of learnable prompts  $P_i^l \in \mathbb{R}^{L \times d}$ , where  $L$  is the number of prompts, for the  $l$ -th self-attention layer as in Figure 2c. Different from previous works [44, 55–57] that employ ViT [11] with global self-attention layers, Swin [33] partitions the input feature map into local windows and performs window-based self-attention [33] to reduce the computation complexity. Thus for the  $l$ -th Swin transformer layer, we duplicate the correlated prompts  $P_i^l$  for each partitioned window and conduct self-attention as

$$\hat{z}^l = \text{MHSA}(\text{CAT}(P_i^l, z^{l-1}))[L:] \quad (2)$$

where MHSA refers to multi-head self-attention and CAT is the concatenation operation.  $z^{l-1}$  is the flattened image feature map within a local window. As MHSA keeps the length of the input sequence, we use  $[L:]$  to preserve only the output corresponding to the input image feature tokens, leaving the spatial shape of the overall image feature maps unchanged for subsequent blocks. By applying the learnable prompts, we jointly train the model in Figure 2b with the detection loss in Equation 1 and the widely used Online-Instance-Matching (OIM) loss [58]  $\mathcal{L}_{oim}$  for each incoming person search domain.

**Diverse attribute matching.** As the domain identity of a test task is unavailable during inference, the learned visual prompts

should be adaptively selected for inference. To tackle this problem, we design a diverse attribute matching module to measure the similarity between an input image and seen domains. Following previous prompt-based continual learning methods [44, 55–57], we duplicate the pre-trained Swin [33] to extract the global visual feature as  $\mathbf{q}_i = F(I_i)$ ,  $\mathbf{q}_i \in \mathbb{R}^c$  of an input image  $I_i$  belonging to domain  $i$ . As the pre-trained model is agnostic to the incoming person search tasks, the visual feature  $\mathbf{q}_i$  implicitly encodes unbiased attributes of domain  $i$ . To capture discriminative domain features, *i.e.* domain attributes, we bind a group of  $N$  learnable embeddings  $K_i^l = \{\mathbf{k}_i^j \in \mathbb{R}^c | j = 1, 2, \dots, N\}$  with the visual prompts  $P_i^l$  as attribute prototypes of domain  $i$ . By maximizing the similarity between  $\mathbf{q}_i$  and  $\mathbf{k}_i^j$  similar to L2P [57], the attribute embeddings are optimized to match the attributes of domain  $i$  during training and the learned prompts can be selected by matching the input image with learned domain attributes during inference. However, this may cause redundancy between different attribute embeddings as the multiple learnable prototypes can easily overfit the same prominent domain attribute. As the domain scenarios can be highly similar between different person search datasets, it also requires learning more diverse domain attributes to mostly uncover the differences between different domains.

For this, we further attach a group of learnable attribute projection embeddings  $W_i^l = \{\mathbf{w}_i^j \in \mathbb{R}^c | j = 1, 2, \dots, N\}$  to  $K_i^l$ .  $W_i^l$  can be regarded as learned channel attention [21] to emphasize a certain attribute of domain  $i$  in an input image. For training on domain  $i$ , given an input scene image  $I_i$ , we calculate the similarity between the image global feature  $\mathbf{q}_i = F(I_i)$  and the attribute prototype  $\mathbf{k}_i^j$  as

$$a_i^j = \mathbf{q}_i \odot \mathbf{w}_i^j \otimes \mathbf{k}_i^j \quad (3)$$

where  $\odot$  denotes channel-wise multiplication and  $\otimes$  stands for calculating cosine similarity. The domain attribute learning loss is thus formulated as

$$\mathcal{L}_{attr} = \sum_{j=1}^N (1 - a_i^j) \quad (4)$$

Besides, both  $\{\mathbf{w}_i^j | j = 1, 2, \dots, N\}$  and  $\{\mathbf{k}_i^j | j = 1, 2, \dots, N\}$  are enforced to be diverse by a diversity loss

$$\mathcal{L}_{div} = \sum_{m=1}^N \sum_{n=1, n \neq m}^N |\mathbf{w}_i^m \otimes \mathbf{w}_i^n|^2 + \sum_{m=1}^N \sum_{n=1, n \neq m}^N |\mathbf{k}_i^m \otimes \mathbf{k}_i^n|^2. \quad (5)$$

The overall training objective on domain  $i$  is thus given by

$$\mathcal{L}_i = \mathcal{L}_{det} + \mathcal{L}_{oim} + \lambda_1 \mathcal{L}_{attr} + \lambda_2 \mathcal{L}_{div} \quad (6)$$

where  $\lambda_1$  and  $\lambda_2$  are predefined loss weights.

During inference, we calculate the similarity between an input image  $I$  and a learned domain as

$$s_{I \rightarrow i} = \frac{1}{N} \sum_{j=1}^N \mathbf{q}_I \odot \mathbf{w}_i^j \otimes \mathbf{k}_i^j \quad (7)$$

where  $\mathbf{q}_I = F(I)$ . Thus the domain index  $d$  of the selected prompts  $P_d^l$  is given by

$$d = \arg \max_i (\{s_{I \rightarrow i} | i = 1, 2, \dots, D\}) \quad (8)$$

where  $D$  is the number of learned domains.

We also note that the learning of diverse domain attributes shares a similar technique with CODA-P [44]. Yet CODA-P is designed for class incremental learning while our work deals with a domain incremental learning problem. CODA-P trains without maximizing the matching scores but implicitly learns attention weights to fuse all prompt components through visual prompt tuning [22]. Yet we explicitly optimize the matching scores to guarantee adaptively selecting of learned prompts. The attention mechanism in CODA-P [44] can also be included in the proposed method to further boost the continual learning performance.

## 4 EXPERIMENTS

### 4.1 Datasets and Evaluation Protocol

**CUHK-SYSU** [58] collected 18,184 images from both movies and real-world street snapshots, presenting 96,143 bounding boxes of pedestrians and 8,432 labeled identities in total. The training set contains 11,206 frames with 5532 identities. The testing set selects 2900 query persons and defines different evaluation protocols with varied gallery sizes.

**PRW** [69] collected scene images from 6 cameras deployed at a campus. In total, it presents 11,816 frames containing 43,110 pedestrian bounding boxes with 932 recognizable identities. The training subset contains 5,134 frames with 432 identities and the test set contains 6,112 frames. Different from CUHK-SYSU [58], the evaluation protocol of PRW takes the full test set as the gallery.

**MovieNet-PS** [40] selected 160K frames of 3,087 identities from 385 movies. The training set keeps persons of 2,087 identities and a test set is with the left 1,000 identities. For training, it presents 3 different settings that preserve at most 10, 30, and 70 instances per identity, resulting in 20K, 54K, and 100K training images in total. For evaluation, the gallery set is constructed in a way similar to [58] with varying sizes.

**Evaluation protocol.** For experiments of CPS, we sequentially train the person search model with the three datasets and then test the model performance on each learned domain without knowing the domain ID. Similar to previous person search research [1, 6, 30, 58, 66], we evaluate the person search performance in the format of **mAP** / **top-1** in the following experiments. The **AP** and **Recall** scores are used to evaluate the model for person detection. Moreover, the gallery size is closely related to the degree of retrieval difficulty in person search. To examine the overall model performance, we calculate the weighted average of a person search performance metric as

$$M_{avg} = \frac{G_c}{G} M_c + \frac{G_p}{G} M_p + \frac{G_m}{G} M_m, G = G_c + G_p + G_m \quad (9)$$

where  $M_c$ ,  $M_p$ , and  $M_m$  are the performance metrics on CUHK-SYSU, PRW, and MovieNet-PS, respectively.  $G_c$ ,  $G_p$  and  $G_m$  are the default gallery size of the respective datasets. We also measure the forgetting of the metrics on learned domains in the same way. Formally, the forgetting measurement is given by the performance decay on domains  $i (i < D)$  after learning on domain  $D$  compared with that when complete learning on domain  $i$ . By default, the continual learning sequence is set to CUHK-SYSU [58]→PRW [69]→MovieNet-PS [40].

### 4.2 Implementation Details

For the compositional person search transformer, we use Swin-S [33] pre-trained on ImageNet-22K [9] as the pre-trained vision transformer. The output size of RoIAlign [20] is set to  $14 \times 14$  to match the size of image features during pre-training. For the pre-training on person detection data, the model is trained for 36 epochs with a batch size of 16. The input image is augmented with random horizontal flip and scaling where the shorter side varies from 480 to 800 during training. We use the AdamW optimizer with an initial learning rate of 0.0001. The learning rate is linearly warmed up during the first 1,000 iterations and decreased by 10 at the  $20^{th}$  epoch. During the continual learning stage, we employ the Adam optimizer with an initial learning rate of 0.0003. For CUHK-SYSU [58] and PRW [69], we keep the image augmentation with batch size 8 and resize the images to  $1333 \times 800$  for evaluation. For MovieNet-PS [40], we instead randomly resize the shorter image side from 160 to 240. The test image size is set to  $720 \times 240$  following [40]. We train the PoPS model for 28 epochs on CUHK-SYSU [58] and PRW [69], and 14 epochs on MovieNet-PS. The loss weights  $\lambda_1$  and  $\lambda_2$  in Equation 6 are both set to 0.1. The test gallery sizes of CUHK-SYSU and MovieNet-PS are 100 and 2000, respectively.

### 4.3 Continual Person Search

To validate the effectiveness of the proposed PoPS method, we conduct continual person search evaluation on *three* types of compared methods: (1) sequentially fine-tuned models including ‘Prompt + FT-seq’ that combines the proposed compositional person search transformer with fixed length prompts, and a representative previous person search model SeqNet [30] with the Swin [33] backbone used in our method; (2) representative continual learning methods applying to the proposed compositional person search transformer; (3) upper-bound model that performs prompt tuning [22] with the compositional person search transformer on the union of all domains. We also test to incorporate the attention mechanism [44] into PoPS. We directly replace each prompt with a prompt component and bind extra attention vectors and keys [44] while the weighted summation of prompt components is restricted in the same domain.

The experimental performances of the aforementioned methods are presented in Table 1. It can be observed that our proposed method obtains the best overall accuracy for CPS. The anti-forgetting performance is also shown to be superior. In addition, incorporating the attention mechanism in CODA-P [44] further improves the accuracy by introducing more learnable parameters. Compared with previous continual learning methods, sequential prompt tuning of the compositional person search transformer performs only slightly inferior on forgetting learned knowledge, while the sequentially fine-tuned SeqNet [30] is marginally behind. We believe the frozen transformer mainly improves the model stability. We also observe that the domain incremental learning methods S-Prompt [55] and our proposed PoPS perform significantly better on anti-forgetting of the PRW [69] domain, demonstrating the superior of domain incremental learning techniques for CPS.

**Table 1: Continual person search performance comparisons between our proposed PoPS and existing methods. We collect both the person retrieval accuracy and forgetting metrics to make a comprehensive understanding of the effectiveness of PoPS. All results are given as mAP / top-1.**

Method	Accuracy ( $\uparrow$ )				Forgetting ( $\downarrow$ )		
	CUHK-SYSU	PRW	MovieNet-PS	Average	CUHK-SYSU	PRW	Average
Prompt + FT-seq	83.2 / 85.3	20.4 / 76.8	38.6 / 84.3	25.6 / 79.7	9.9 / 8.8	25.7 / 7.6	25.4 / 7.6
SeqNet [30] (Swin [33])-seq	75.6 / 77.3	19.9 / 75.6	42.1 / 87.2	26.3 / 79.4	17.8 / 16.8	25.9 / 9.1	25.8 / 9.2
L2P [57]	85.0 / 87.1	21.7 / 77.0	36.4 / 82.6	26.4 / 79.4	7.8 / 6.7	24.3 / 7.1	24.0 / 7.1
DualPrompt [56]	81.7 / 83.0	18.1 / 71.5	36.8 / 83.2	23.7 / 75.4	9.3 / 8.7	23.6 / 7.8	23.4 / 7.8
CODA-P [44]	85.9 / 86.7	24.7 / 77.5	<b>40.2 / 85.7</b>	29.7 / 80.6	9.7 / 9.0	26.8 / 9.3	26.5 / 9.3
S-Prompt [55]	81.3 / 83.6	16.7 / 69.8	32.0 / 79.4	21.5 / 73.2	7.2 / 6.3	6.3 / 2.6	6.3 / 2.7
<b>PoPS</b>	<u>86.3 / 87.3</u>	<u>42.8 / 83.3</u>	35.4 / 82.3	<u>42.0 / 84.1</u>	<b>5.8 / 6.0</b>	<b>0.1 / 0.2</b>	<b>0.2 / 0.3</b>
<b>PoPS + Attention [44]</b>	<b>87.5 / 88.6</b>	<b>49.6 / 85.8</b>	<u>39.7 / 85.9</u>	<b>48.2 / 86.9</b>	<u>6.8 / 6.4</u>	<u>0.2 / 0.2</u>	<u>0.3 / 0.3</u>
Prompt + upper-bound	91.6 / 92.8	46.3 / 84.1	40.1 / 86.4	45.9 / 85.8	-	-	-

### 4.4 Analytical Studies

In this subsection, we conduct ablation experiments to understand the impact of the designed modules in PoPS for CPS. For the convenience of notation, we use MVN as the short for **MovieNet-PS**.

**The effectiveness of detection pre-training.** To construct an effectively pre-trained person search transformer without exhaustive pre-training from scratch, we design a compositional person search transformer by expanding a pre-trained Swin with a Simple Feature Pyramid. The added sub-network is then pre-trained by a person detection task. To examine the effectiveness of detection pre-training, we test the pre-trained model on the three person search datasets for person detection and collect the results in Table 2. We also test a standard detector Faster R-CNN [41] independently trained on the datasets for comparisons. Although the pre-trained PoPS is evaluated in a cross-domain manner, it can be observed the person detection result is still acceptable compared with the fully supervised Faster R-CNN [41]. This suggests that the detection pre-training is effective for preparing a person search transformer for prompt-based continual learning.

**Table 2: Person detection performance comparisons between the pre-trained compositional person search transformer and a standard Faster R-CNN [41] detector.**

Method	CUHK-SYSU	PRW	MVN
	AP/Recall	AP/Recall	AP/Recall
PoPS	81.8/88.0	90.0/95.4	73.7/83.3
Faster R-CNN [41]	87.0/92.9	93.0/96.1	89.4/96.9

**The effect of attribute projection.** To encourage diverse domain attribute learning for adaptive prompt selection, we introduce the attribute projection embeddings  $W_i^l$  to reveal the domain attribute information in an input image before matching it with the learned attribute prototypes. To validate the effect of the attribute projection embeddings, we test PoPS with and without them as in Table 3. Without the attribute projection embeddings, we directly maximize the similarity between the image feature  $q_i$  and its

correlated diverse attribute prototypes  $K_i^l$  during training. It can be observed that introducing the attribute projection consistently improves model performance on early domains, suggesting that this mechanism enables more robust prompt selection.

**Table 3: Continual person search performance comparisons between PoPS with and without the attribute projection.**

Proj.	CUHK-SYSU	PRW	MVN	Average
×	81.3 / 82.3	40.9 / 82.0	35.5 / 82.1	40.6 / 83.0
✓	86.3 / 87.3	42.8 / 83.3	35.4 / 82.3	42.0 / 84.1

#### Comparison between deep prompt and shallow prompt.

As the vision transformers [11, 33] are formulated in a multi-layer architecture, it is also important to explore which layers to insert the prompts. For this, VPT [22] test both shallow (prompts for only the first layer) and deep (prompts for every layer) prompts with ViT [11]. Different from ViT [11], Swin [33] is composed of 4 stages where each of them contains several layers. We therefore test shallow and deep prompts upon stages for PoPS as in Table 4. The results suggest that the model requires deep stage-wise prompts to achieve the best performance. Moreover, adding prompts to the last stage gives a significant improvement in the performance. This is mainly due to that ‘Stage4’ processes the features in a different way, *i.e.* instance-wise, thus requiring extra prompts to adapt well to the person search task.

**Table 4: Continual person search performance of PoPS when prompting different Swin [33] stages.**

Stages	CUHK-SYSU	PRW	MVN	Average
(1,)	67.6 / 69.4	33.6 / 65.3	29.9 / 65.6	33.5 / 66.2
(1,2)	69.1 / 70.6	34.3 / 66.7	30.3 / 66.9	34.2 / 67.6
(1,2,3)	72.6 / 73.8	36.1 / 70.0	31.3 / 69.9	35.8 / 70.9
(1,2,3,4)	86.3 / 87.3	42.8 / 83.3	35.4 / 82.3	42.0 / 84.1

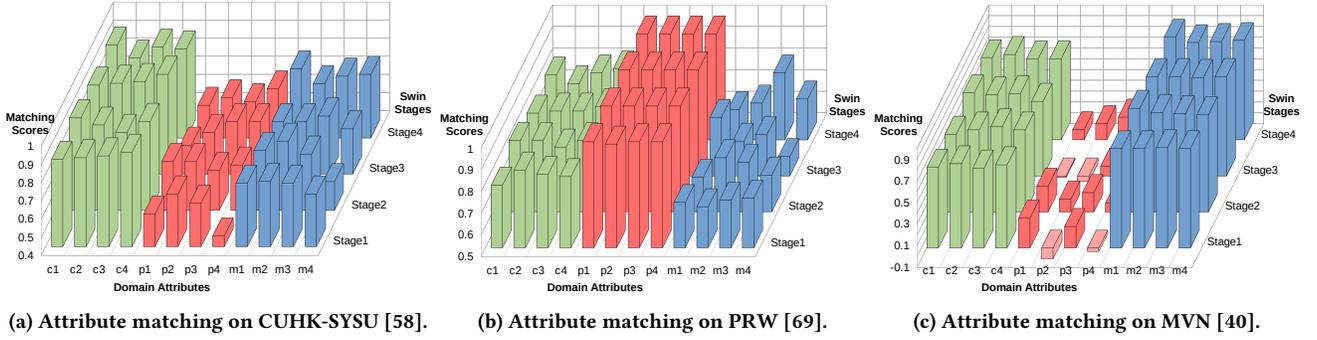


Figure 3: Exampler visualization of prompt selection on CUHK-SYSU [58], PRW [69] and MVN [40]. We denote by  $c_i$ ,  $p_i$  and  $m_i$  the learned  $i$ -th domain attribute from the three datasets, respectively.

**The number of prompts.** With the person search transformer frozen during training, the number of learnable prompts for each transformer layer is also important. To validate the effect of different numbers of learnable prompts, we conduct the experiments in Table 5. By default, we set the number of prompts  $L$  to 16. It can be observed that decreasing the prompts leads to a significant drop in model performance while adding more prompts only slightly improves the results. As the window attention performs self-attention [49] on relatively short image feature sequences (e.g.  $7 \times 7$ ), we assume that the inserted prompts should be well balanced with the image feature. Thus the optimal number of prompts should be moderate.

Table 5: Continual person search performance of PoPS with different numbers of learnable prompts.

$L$	CUHK-SYSU	PRW	MVN	Average
4	81.3 / 82.3	39.9 / 82.1	33.6 / 80.3	39.3 / 82.7
8	84.7 / 86.0	40.7 / 82.1	34.1 / 80.9	30.1 / 82.9
16	86.3 / 87.3	42.8 / 83.3	35.4 / 82.3	42.0 / 84.1
32	86.5 / 87.8	43.3 / 83.6	36.0 / 82.5	42.6 / 84.4

**The number of learnable domain attributes.** To properly select learned prompts for CPS, we learn to capture  $N$  diverse attributes for recognizing each domain. To validate the impact of  $N$ , we test the CPS performance of PoPS with varied  $N$  as in Table 6. It can be observed that learning 4 diverse domain attributes results in better overall performance. When setting a small  $N$ , the model may fail to capture sufficient distinct domain features. And the attribute embeddings may overfit on training samples when setting a large  $N$  and impedes the person search performance. We thus set  $N = 4$  for each domain by default.

**Visualization of prompt selection.** To qualitatively understand the effectiveness of adaptive prompt selection with diverse attribute matching, we visualize the matching scores between input images and learned attributes across different domains. The scores are averaged on 10 randomly selected test images from the learned domains as in Figure 3. We also average the scores in different layers within the same stage to obtain stage-level matching scores. It can be observed that the matching scores between the test image

Table 6: Evaluation of continual person search for PoPS with different numbers of learnable domain attributes.

$N$	CUHK-SYSU	PRW	MVN	Average
2	85.7 / 86.9	42.5 / 82.4	32.1 / 79.1	41.0 / 82.6
4	86.3 / 87.3	42.8 / 83.3	35.4 / 82.3	42.0 / 84.1
8	84.9 / 86.3	40.7 / 81.9	32.4 / 80.0	39.7 / 82.5

and the truly corresponding domain attributes are relatively high compared with the distractors, suggesting the effectiveness of diverse attribute matching. As the learning domains can be highly similar in CPS, we also observe that there always exists a highly similar distracting domain when testing the model.

## 5 CONCLUSION

This work introduces a new challenging yet practical CPS task that learns from sequentially incoming domains and adaptively completes the person search task for any learned domain. To better balance the stability and plasticity of the model to consistently adapt to new domains without catastrophic forgetting of seen domains, we design a Prompt-based Continual Person Search model (PoPS). To reduce the cost of pre-training a person search transformer from scratch, we first propose a compositional person search transformer that expands a pre-trained hierarchical vision transformer with a Simple Feature Pyramid to enable person localization. We then pre-train the Simple Feature Pyramid with only a moderate number of person detection data to construct a fully pre-trained person search transformer. On top of that, we design a domain incremental prompt pool with a diverse attribute matching module. During training, the prompts are independently learned to encode the domain-oriented knowledge, and a group of paired attribute projections and prototype embeddings are forced to diversely capture distinct domain attributes. This facilitates the adaptive selection of learned prompts by matching an input image with the learned attributes across domains for model inference. For future works, we shall explore designing a more efficient continual learning framework and collecting more realistic domains to better tackle the CPS problem.

## ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China 62276016 and 62372029.

## REFERENCES

- [1] Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, and Fahad Shahbaz Khan. 2022. Pstr: End-to-end one-step person search with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9458–9467.
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.
- [3] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2018. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420* (2018).
- [4] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. 2020. Hierarchical Online Instance Matching for Person Search. In *AAAI*.
- [5] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. 2020. Person search by separated modeling and a mask-guided two-stream cnn model. *IEEE Transactions on Image Processing* 29 (2020), 4669–4682.
- [6] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. 2020. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12615–12624.
- [7] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. 2023. Beyond Appearance: a Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [8] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9640–9649.
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [10] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. 2020. Instance guided proposal network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2585–2594.
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
- [12] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6824–6835.
- [13] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. 2021. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14750–14759.
- [14] Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. 2023. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11483–11493.
- [15] Wenhao Ge, Junlong Du, Ancong Wu, Yuqiao Xian, Ke Yan, Feiyue Huang, and Wei-Shi Zheng. 2022. Lifelong person re-identification by pseudo task knowledge preservation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 688–696.
- [16] Chuchu Han, Kai Su, Dongdong Yu, Zehuan Yuan, Changxin Gao, Nong Sang, Yi Yang, and Changhu Wang. 2021. Weakly supervised person search with region siamese networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12006–12015.
- [17] Chuchu Han, Kai Su, Dongdong Yu, Zehuan Yuan, Changxin Gao, Nong Sang, Yi Yang, and Changhu Wang. 2021. Weakly Supervised Person Search with Region Siamese Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12006–12015.
- [18] Chuchu Han, Zhedong Zheng, Changxin Gao, Nong Sang, and Yi Yang. 2021. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1505–1512.
- [19] Tyler L Hayes, Nathan D Cahill, and Christopher Kanan. 2019. Memory efficient experience replay for streaming learning. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 9769–9776.
- [20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.
- [21] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.
- [22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*. Springer, 709–727.
- [23] Dahui Jung, Dongyoon Han, Jihwan Bang, and Hwanjun Song. 2023. Generating Instance-level Prompts for Rehearsal-free Continual Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11847–11857.
- [24] Dahui Jung, Dongyoon Han, Jihwan Bang, and Hwanjun Song. 2023. Generating Instance-level Prompts for Rehearsal-free Continual Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11847–11857.
- [25] Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Didier Stricker, Federico Tombari, and Muhammad Zeshan Afzal. 2023. Introducing language guidance in prompt-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11463–11473.
- [26] Xu Lan, Xiatian Zhu, and Shaogang Gong. 2018. Person search by multi-scale matching. In *Proceedings of the European conference on computer vision (ECCV)*. 536–552.
- [27] Junjie Li, Yichao Yan, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. 2022. Domain adaptive person search. In *European Conference on Computer Vision*. Springer, 302–318.
- [28] Junjie Li, Yichao Yan, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. 2022. Domain adaptive person search. In *Proceedings of the European Conference on Computer Vision*. Springer, 302–318.
- [29] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. 2022. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*. Springer, 280–296.
- [30] Zhengjia Li and Duoqian Miao. 2021. Sequential end-to-end network for efficient person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2011–2019.
- [31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 740–755.
- [32] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. 2020. Dual context-aware refinement network for person search. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3450–3459.
- [33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.
- [34] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. 2019. Query-guided end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 811–820.
- [35] Minyoung Oh, Duhyun Kim, and Jae-Young Sim. 2024. Domain Generalizable Person Search Using Unreal Dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4361–4368.
- [36] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1406–1415.
- [37] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. 2021. Lifelong person re-identification via adaptive knowledge accumulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7901–7910.
- [38] Nan Pu, Yu Liu, Wei Chen, Erwin M Bakker, and Michael S Lew. 2022. Meta reconciliation normalization for lifelong person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*. 541–549.
- [39] Nan Pu, Zhun Zhong, Nicu Sebe, and Michael S Lew. 2023. A memorizing and generalizing framework for lifelong person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [40] Jie Qin, Peng Zheng, Yichao Yan, Rong Quan, Xiaogang Cheng, and Bingbing Ni. 2023. MovieNet-PS: a large-scale person search dataset in the wild. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
- [42] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. 2018. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123* (2018).
- [43] Bing Shuai, Xinyu Li, Kaustav Kundu, and Joseph Tighe. 2022. Id-free person similarity learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14689–14699.
- [44] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. 2023. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11909–11919.

- [45] Zhicheng Sun and Yadong Mu. 2022. Patch-based knowledge distillation for lifelong person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*. 696–707.
- [46] Yu-Ming Tang, Yi-Xing Peng, and Wei-Shi Zheng. 2023. When prompt-based incremental learning does not meet strong pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1706–1716.
- [47] Yanling Tian, Di Chen, Yunan Liu, Shanshan Zhang, and Jian Yang. 2022. Grouped Adaptive Loss Weighting for Person Search. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6774–6782.
- [48] Gido M Van de Ven and Andreas S Tolias. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734* (2019).
- [49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [50] Benzhi Wang, Yang Yang, Jinlin Wu, Guo-jun Qi, and Zhen Lei. 2023. Self-similarity driven scale-invariant learning for weakly supervised person search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1813–1822.
- [51] Benzhi Wang, Yang Yang, Jinlin Wu, Guo-jun Qi, and Zhen Lei. 2023. Self-similarity Driven Scale-invariant Learning for Weakly Supervised Person Search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1813–1822.
- [52] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. 2020. Tcts: A task-consistent two-stage framework for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11952–11961.
- [53] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. 2024. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *Advances in Neural Information Processing Systems* 36 (2024).
- [54] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [55] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. 2022. S-prompts learning with pre-trained transformers: An occam’s razor for domain incremental learning. *Advances in Neural Information Processing Systems* 35 (2022), 5682–5695.
- [56] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. 2022. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*. Springer, 631–648.
- [57] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 139–149.
- [58] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3415–3424.
- [59] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. 2014. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the 22nd ACM International Conference on Multimedia*. 937–940.
- [60] Yichao Yan, Jimpeng Li, Shengcai Liao, Jie Qin, Bingbing Ni, Ke Lu, and Xiaokang Yang. 2022. Exploring visual context for weakly supervised person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3027–3035.
- [61] Yichao Yan, Jimpeng Li, Shengcai Liao, Jie Qin, Bingbing Ni, Ke Lu, and Xiaokang Yang. 2022. Exploring visual context for weakly supervised person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3027–3035.
- [62] Yichao Yan, Jimpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. 2021. Anchor-free person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7690–7699.
- [63] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. 2019. Learning context graph for person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2158–2167.
- [64] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence* 44, 6 (2021), 2872–2893.
- [65] Chunlin Yu, Ye Shi, Zimo Liu, Shenghua Gao, and Jingya Wang. 2023. Life-long person re-identification via knowledge refreshing and consolidation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 3295–3303.
- [66] Rui Yu, Dawei Du, Rodney LaLonde, Daniel Davila, Christopher Funk, Anthony Hoogs, and Brian Clipp. 2022. Cascade transformers for end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7267–7276.
- [67] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. 2021. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 516–525.
- [68] Wei Zhang, Lingxiao He, Peng Chen, Xingyu Liao, Wu Liu, Qi Li, and Zhenan Sun. 2021. Boosting End-to-end Multi-Object Tracking and Person Search via Knowledge Distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1192–1201.
- [69] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1367–1376.
- [70] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.

## SUPPLEMENTARY EXPERIMENTS

**Table 7: Person detection performance comparisons between different pretraining methods.**

Pretrain	CUHK-SYSU		PRW		MVN	
	AP	Recall	AP	Recall	AP	Recall
ImageNet-22k [9]	81.8	88.0	90.0	95.4	73.7	83.3
ImageNet-1k [9]	61.9	69.0	65.6	73.8	59.5	80.9
SOLIDER [7]	82.1	88.3	89.3	95.5	73.9	85.3

**Table 8: Continual person search performance comparisons between different pretraining methods. We use IMG-22k as the short for ImageNet-22k for illustration purposes.**

Pretrain	CUHK-SYSU	PRW	MVN	Average
IMG-22k [9]	86.3 / 87.3	42.8 / 83.3	35.4 / 82.3	42.0 / 84.1
SOLIDER [7]	89.9 / 90.9	29.5 / 76.5	21.0 / 60.5	28.2 / 72.8

**The effect of vision transformer pre-training.** Following previous peompt-based continual learning methods [44, 55–57], we employ the Swin Transformer [33] pre-trained on ImageNet-22k [9] to gaurentee the generalization capaility of the pre-trained transformer. To validate the impact of the pre-training, we further test to employ Swin pre-trained with ImageNet-1K [9] in PoPS. We also note that recent work, SOLIDER [7], also presents an effective Swin variant pre-trained on large scale unlabeled person images [13] and achieves superior performances when fine-tuned on downstream tasks. For this, we also test PoPS based on the pre-trained SOLIDER for continual person search.

As is shown in Table 7, we first compare the pre-training methods on the person detection pre-training stage. It can be observed that although SOLIDER [7] is trained on large scale person images, the performance on person detection pre-training is similar to the ImageNet-22k pre-trained Swin Transformer. In contrast, the ImageNet-1k pre-trained model falls largely behind, suggesting that the scale of pre-training data is significant to enable effective prompt-based learning. We further conduct continual person search with the SOLIDER [7] pre-trained model as in Table 8. Although the model with SOLIDER performs more robustly on the CUHK-SYSU dataset, the model fails to fit the more challenging PRW and MovieNet-PS datasets. As the scene images usually contain complex background objects in person search, we hypothesize that the Swin trained with only person images can be less robust than the ImageNet-22k pre-trained version, especially on challenging person search datasets.

**Data efficiency of the compositional person search transformer.** To reduce the cost of pretraining a person search transformer, we intriduce a compositional person search transformer by extending a pretrained vision transformer with a detection sub-network. We then only optimize the detection sub-network by the person detection task to form a pretrained person search transformer. The detection pretrain task reduce the cost of collecting

**Table 9: Model performance in low-data environments.**

Data	CUHK-SYSU	PRW	MVN	Average
25%	85.7 / 86.5	42.0 / 82.1	33.1 / 80.3	40.4 / 81.7
50%	86.4 / 87.5	41.3 / 82.9	34.6 / 82.0	40.2 / 82.7
100%	86.3 / 87.3	42.8 / 83.3	35.4 / 82.3	42.0 / 84.1

data, and the limited learnable parameters can be optimized with less data. To quantitatively understand the data efficiency of the transformer, we test to pretrain the model with reduced data and then conduct the continual learning process. As shown in Table 9, largely reducing the amount of pretrain data to 50% or even 25% only slightly hinders the continual person search performance, which demonstrates the data efficiency of compositional person search transformer.

**Comparison between prepend and prefix prompt tuning.** To enable effective learning of visual prompts, DualPrompt [56] explores conducting prefix prompt tuning instead of directly prepending visual prompts and obtains improved model performance. CODA-P [44] also follows the prefix prompt tuning mechanism. As we employ a different vision transformer (Swin [33] vs ViT [11]) from those works, we conduct comparisons between prepend and prefix prompt tuning in the proposed PoPS. However, as in Table 10, it can be observed that changing the prompt tuning mechanism barely improves the model performance. For simplicity, we thus evaluate all compared methods with the prepend prompt tuning mechanism.

**Table 10: Continual person search performance comparisons between different prompt tuning methods.**

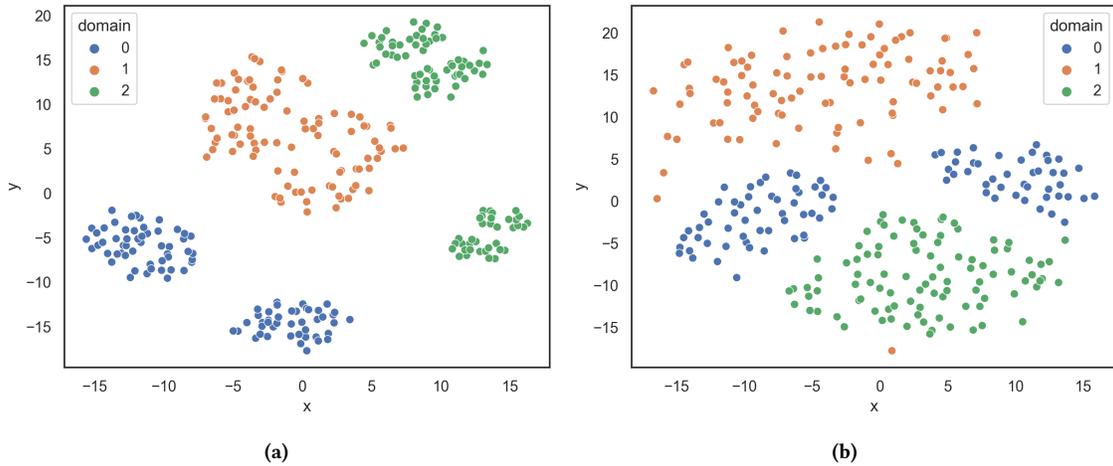
Pretrain	CUHK-SYSU	PRW	MVN	Average
prepend	86.3 / 87.3	42.8 / 83.3	35.4 / 82.3	42.0 / 84.1
prefix	85.2 / 86.4	43.2 / 83.8	34.8 / 81.9	41.7 / 83.4

**Distribution of learned domain attributes.** To qualitatively understand the correlation between learned domain attributes across different person search domains, we conduct t-sne visualization of the learned domain attribute prototypes as well as the attribute projection embeddings as in Figure 4a and Figure 4b. We refer to CUHK-SYSU [58], PRW [69], and MovieNet-PS [40] as domain 0, 1, and 2, respectively. It can be observed that the learned attribute prototypes effectively capture the distinct differences between learned domains. The attribute projection embeddings also clearly reflect the boundary between different domains, demonstrating the effectiveness of the proposed method.

**Person detection performance in continual person search.** Person search is a multi-task learning problem that jointly learns to conduct person detection and re-identification [58]. In addition to the evaluated person search performance, the person detection capability also has an impact to the overall person search accuracy and is affected during the continual learning procedure. We thus evaluate the person detection performance of the proposed method and the compared methods to make a more comprehensive understanding. Different from the evaluation of person retrieval performances, the person detection performances are tested on

**Table 11: Continual person detection performance of our proposed PoPS. We collect both the person detection accuracy and forgetting metrics to make a comprehensive understanding of the effectiveness of PoPS. All results are given as AP / Recall.**

Method	Accuracy ( $\uparrow$ )				Forgetting ( $\downarrow$ )		
	CUHK-SYSU	PRW	MovieNet-PS	Average	CUHK-SYSU	PRW	Average
Pre-trained PoPS	81.8 / 88.0	90.0 / 95.4	73.7 / 83.3	81.8 / 88.9	-	-	-
Prompt + FT-seq	72.5 / 78.7	88.3 / 92.4	85.6 / 95.4	82.1 / 88.8	13.5 / 12.5	5.0 / 5.1	9.3 / 8.8
L2P [57]	76.2 / 82.8	89.5 / 94.3	<u>85.4 / 94.9</u>	83.9 / 90.7	10.0 / 8.5	3.5 / 3.1	6.8 / 5.8
DualPrompt [56]	79.6 / 85.2	90.3 / 94.8	85.1 / 94.4	85.0 / 91.5	6.5 / 5.9	2.4 / 2.4	4.4 / 4.1
CODA-P [44]	80.2 / 87.2	88.7 / 95.7	<b>85.6 / 95.2</b>	84.8 / 92.7	7.0 / 4.8	5.7 / 2.0	6.4 / 3.4
S-Prompt [55]	83.5 / 87.9	89.4 / 94.4	84.8 / 94.1	85.9 / 92.1	3.0 / 4.1	2.1 / 2.4	2.6 / 3.3
<b>PoPS</b>	<u>85.0 / 90.4</u>	<u>92.5 / 97.2</u>	84.3 / 94.0	<u>87.3 / 93.9</u>	1.0 / 0.6	<b>0.1 / 0.1</b>	<u>0.6 / 0.4</u>
<b>PoPS + Attention [44]</b>	<b>85.6 / 90.6</b>	<b>93.6 / 97.5</b>	84.5 / 94.6	<b>87.9 / 94.2</b>	<b>0.9 / 0.7</b>	<b>0.1 / 0.1</b>	<b>0.5 / 0.4</b>
Prompt + upper-bound	83.9 / 89.9	92.1 / 97.0	89.3 / 94.8	88.4 / 93.9	-	-	-

**Figure 4: (a) T-sne visualization of learned domain attribute prototypes in PoPS. (b) T-sne visualization of learned domain attribute projections in PoPS.**

approximately equal-sized test sets, we thus directly average the results on different domains to obtain an overall performance measurement.

As shown in Table 11, our proposed PoPS consistently achieves superior overall person detection accuracy compared with previous prompt-based continual learning methods [44, 55–57]. The anti-forgetting performance is also outstanding on both CUHK-SYSU [58] and PRW [69] datasets. It is also worth noting that the overall person detection performance is less hindered by the continual

learning procedure compared with the person search performance. This is mainly due to the person detection sub-tasks sharing more common knowledge between different domains while the person retrieval sub-task requires more sophisticated domain-specific knowledge. We also observe that PoPS even performs better than the jointly trained upper-bound. This is mainly caused by the annotation bias in different datasets, e.g. some of the small background persons are not annotated in CUHK-SYSU [58] but annotated in other datasets, which may confuse the model during training.