

Supplementary Materials: Prompting Continual Person Search

Anonymous Authors

1 SUPPLEMENTARY EXPERIMENTS

Table 1: Person detection performance comparisons between different pretraining methods.

Pretrain	CUHK-SYSU		PRW		MVN	
	AP	Recall	AP	Recall	AP	Recall
ImageNet-22k [2]	81.8	88.0	90.0	95.4	73.7	83.3
ImageNet-1k [2]	61.9	69.0	65.6	73.8	59.5	80.9
SOLIDER [1]	82.1	88.3	89.3	95.5	73.9	85.3

Table 2: 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 [2]	86.3 / 87.3	42.8 / 83.3	25.4 / 77.3	39.1 / 81.9
SOLIDER [1]	89.9 / 90.9	29.5 / 76.5	11.0 / 55.5	25.7 / 71.6

The effect of vision transformer pre-training. Following previous prompt-based continual learning methods [7–10], we employ the Swin Transformer [5] pre-trained on ImageNet-22k [2] to guarantee the generalization capability of the pre-trained transformer. To validate the impact of the pre-training, we further test to employ Swin pre-trained with ImageNet-1K [2] in PoPS. We also note that recent work, SOLIDER [1], also presents an effective Swin variant pre-trained on large scale unlabeled person images [4] 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 1, we first compare the pre-training methods on the person detection pre-training stage. It can be observed that although SOLIDER [1] 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 [1] pre-trained model as in Table 2. 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.

Comparison between prepend and prefix prompt tuning.

To enable effective learning of visual prompts, DualPrompt [9] explores conducting prefix prompt tuning instead of directly prepending visual prompts and obtains improved model performance. CODA-P [7] also follows the prefix prompt tuning mechanism. As we employ a different vision transformer (Swin [5] vs ViT [3]) from those works, we conduct comparisons between prepend and prefix prompt tuning in the proposed PoPS. However, as in Table 3, 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 3: 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	25.4 / 77.3	39.1 / 81.9
prefix	85.2 / 86.4	43.2 / 83.8	24.8 / 75.9	39.2 / 81.9

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

As shown in Table 4, our proposed PoPS consistently achieves superior overall person detection accuracy compared with previous prompt-based continual learning methods [7–10]. The anti-forgetting performance is also outstanding on both CUHK-SYSU [11] and PRW [12] 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 [11] but annotated in other datasets, which may confuse the model during training.

2 EXTENSIVE VISUALIZATION

Distribution of learned domain attributes. To qualitatively understand the correlation between learned domain attributes across

Table 4: 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 [10]	76.2 / 82.8	89.5 / 94.3	85.4 / 94.9	83.9 / 90.7	10.0 / 8.5	3.5 / 3.1	6.8 / 5.8
DualPrompt [9]	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 [7]	80.2 / 87.2	88.7 / 95.7	85.6 / 95.2	84.8 / 92.7	7.0 / 4.8	5.7 / 2.0	6.4 / 3.4
S-Prompt [8]	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
PoPS	<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	0.1 / 0.1	<u>0.6 / 0.4</u>
PoPS + Attention [7]	85.6 / 90.6	93.6 / 97.5	84.5 / 94.6	87.9 / 94.2	0.9 / 0.7	0.1 / 0.1	0.5 / 0.4
Prompt + upper-bound	83.9 / 89.9	92.1 / 97.0	89.3 / 94.8	88.4 / 93.9	-	-	- -

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 1 and figure 2. We refer to CUHK-SYSU [11], PRW [12], and MovieNet-PS [6] 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.

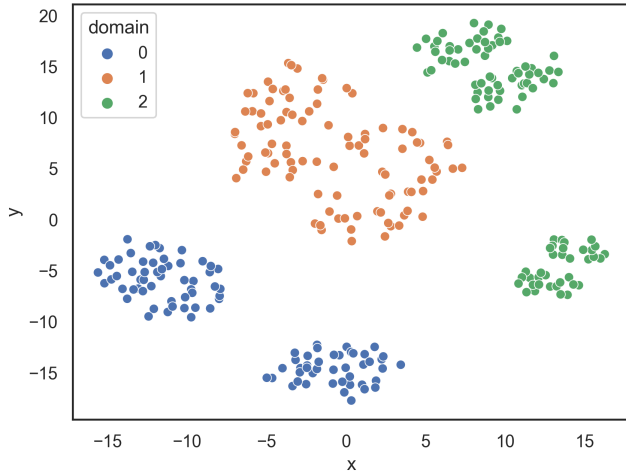


Figure 1: T-sne visualization of learned domain attribute prototypes in PoPS.

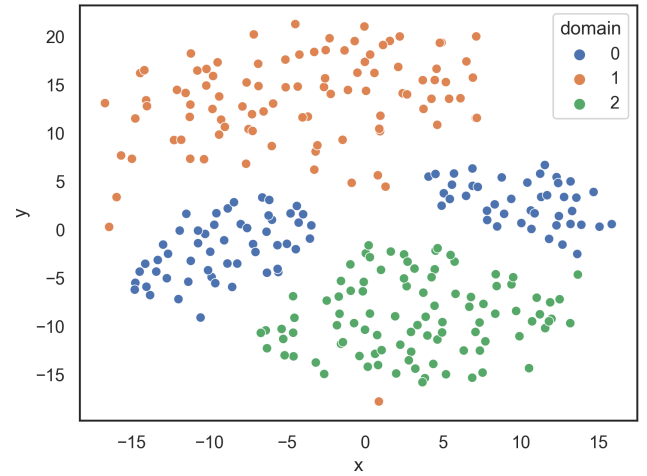


Figure 2: T-sne visualization of learned domain attribute projections in PoPS.

REFERENCES

- [1] 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*.
- [2] 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.
- [3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua 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*.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.

- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.

233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348