Supplementary Material for
SLCA: Slow Learner with Classifier Alignment for Continual Learning on a Pre-trained Model

Gengwei Zhang[1]*, Liyuan Wang[2]*, Guoliang Kang[3,4], Ling Chen[1], Yunchao Wei[5,6†]
[1] University of Technology Sydney [2] Tsinghua University
[3] Beihang University [4] Zhongguancun Laboratory
[5] Institute of Information Science, Beijing Jiaotong University
[6] Beijing Key Laboratory of Advanced Information Science and Network
{zgwdavid, kgl.prml, wychao1987}@gmail.com; wly19@mail.tsinghua.org.cn; ling.chen@uts.edu.au

1. More Details and Results.

Implementation Details. All baselines follow an implementation similar to the one described in [4, 3]. Specifically, we adopt a pre-trained ViT-B/16 backbone. We use an Adam optimizer for prompting-based approaches that keep the representation layer fixed, while a SGD optimizer for other baselines that update the entire model, with the same batch size of 128. The original implementation of [4, 3] adopts a constant learning rate of 0.005 for all baselines, while our slow learner using 0.0001 for the representation layer and 0.01 for the classification layer. In practice, we observe that supervised pre-training usually converges faster than self-supervised pre-training in downstream continual learning. Therefore, for supervised pre-training, we train all baselines for 20 epochs on Split CIFAR-100 and 50 epochs on other benchmarks. For self-supervised pre-training, we train all baselines for 90 epochs on all benchmarks.

Extended Analysis. In this section, we provide extended results to support the main claims in our paper. First, we present the CKA similarity of pre-trained representation (1) before and after learning downstream tasks in Fig. 1, and (2) after joint training and after continual learning in Fig. 2.

Results on Additional Dataset. Except CIFAR-100, CUB-200-2011, ImageNet-R and Cars-196, we further consider a subset of DomainNet with 345-class sketch images (for short, Sketch-345). Our SLCA delivers consistently strong performance as shown in Table 1.

Combine with other methods. In the main text, the efficacy of SL has been widely validated by combining it with all baseline methods. We have further validated the efficacy of CA, presenting representative non-replay and replay methods on IN21K-Sup as shown in Table 2.

Method	Sketch-345, IN21K-Sup Last-Acc (%)	Sketch-345, IN21K-Sup Inc-Acc (%)	Sketch-345, IN1K-Self Last-Acc (%)	Sketch-345, IN1K-Self Inc-Acc (%)
Joint-Training	$72.18_{\pm 0.03}$	-	$66.04_{\pm 0.07}$	-
Seq FT	$40.40_{\pm 14.87}$	$46.91_{\pm 24.25}$	$12.98_{\pm 4.09}$	$38.80_{\pm 5.49}$
w/ SL	$63.41_{\pm 0.53}$	$71.24_{\pm 0.67}$	$56.94_{\pm 0.05}$	$66.07_{\pm 0.38}$
w/ SL+CA	$\mathbf{64.92}_{\pm 0.81}$	$\mathbf{72.69}_{\pm 0.57}$	$\mathbf{59.88}_{\pm 0.06}$	$\mathbf{67.99}_{\pm 0.54}$

Table 1. Results on Sketch-345, a subset of DomainNet dataset [2].

Method	CIFAR-100	ImageNet-R	CUB-200	Cars-196
EWC	$47.01_{\pm 0.29}$	$35.00_{\pm 0.43}$	$51.28_{\pm 2.37}$	$47.02_{\pm 3.90}$
EWC w/ SL	$89.30_{\pm 0.23}$	$70.27_{\pm 1.99}$	$81.62_{\pm 0.34}$	$64.50_{\pm 0.36}$
EWC w/ SL+CA	$\mathbf{90.61}_{\pm 0.17}$	$\mathbf{71.48}_{\pm 0.31}$	$\mathbf{84.29}_{\pm 0.37}$	$\mathbf{69.61}_{\pm 0.29}$
BiC	$66.11_{\pm 1.76}$	$52.14_{\pm 1.08}$	$78.69_{\pm 1.97}$	$55.03_{\pm 3.27}$
BiC w/ SL	$88.45_{\pm 0.57}$	$64.89_{\pm 0.80}$	$81.91_{\pm 2.59}$	$63.10_{\pm 5.71}$
BiC w/ SL+CA	$\mathbf{91.57}_{\pm 0.13}$	$\mathbf{74.49}_{\pm 0.08}$	$\mathbf{86.82}_{\pm 0.69}$	$\mathbf{73.90}_{\pm 0.38}$

Table 2. Ablations for CA combining with EWC and BiC.

References

[1] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[2] Xingchao Peng et al. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

[3] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. *arXiv preprint arXiv:2204.04799*, 2022.

[4] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, et al. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

*Equal contribution.
†Corresponding author.

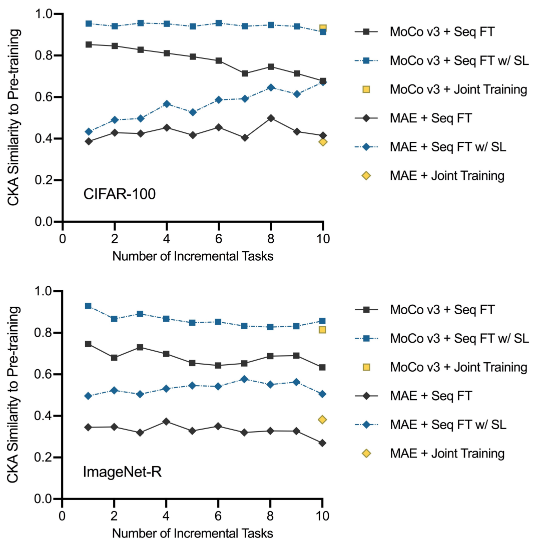

Figure 1. CKA similarity of pre-trained representations before and after learning downstream tasks.

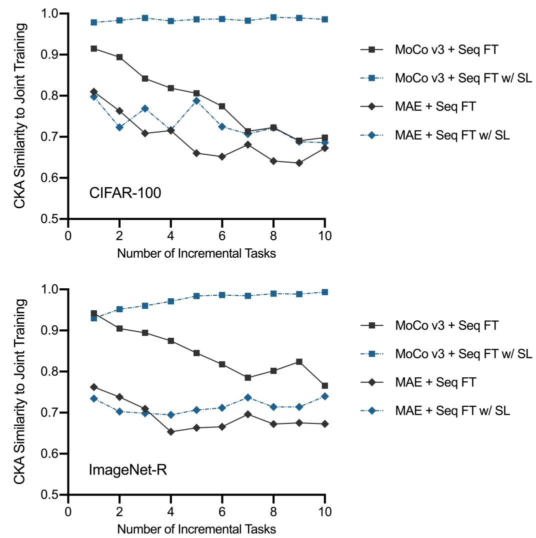

Figure 2. CKA similarity of pre-trained representations after joint training and after continual learning.

Benchmark	Pre-trained	0.005†	0.001	0.0001	0.00001	0.000001	Fixed θ_{rps}
Split CIFAR-100	IN21K-Sup	44.77 ±13.8	83.04±1.46	88.86±0.83	88.81±0.46	85.11±0.42	63.75±0.67
Split ImageNet-R	IN21K-Sup	26.95 ±11.8	70.38±0.80	71.80±1.45	62.64±2.35	53.57±4.33	34.64±14.3
Split CUB-200	IN21K-Sup	40.02 ±1.08	60.02±1.24	68.07±1.09	66.58±3.93	64.38±3.36	60.44±1.80
Split Cars-196	IN21K-Sup	27.57 ±1.79	15.74±26.3	49.74±1.25	30.66±9.01	24.85±7.90	24.51±6.90
Split CIFAR-100	IN1K-Self	27.99 ±5.16	81.49±0.75	81.47±0.55	81.57±0.14	78.61±0.29	77.30±0.56
Split ImageNet-R	IN1K-Self	45.84 ±4.19	68.72±0.48	64.43±0.44	59.19±0.33	54.54±0.32	51.97±0.17
Split CUB-200	IN1K-Self	45.35 ±1.38	68.58±1.16	61.67±1.37	56.46±1.86	55.10±2.13	55.54±1.55
Split Cars-196	IN1K-Self	35.96 ±2.04	58.39±2.31	52.91±1.61	43.64±0.73	41.74±0.23	43.16±0.12

Table 3. Continual learning performance with different learning rates of the representation layer. Here we present the Last-Acc (%) after continual learning of all classes. IN21K-Sup: supervised pre-training on ImageNet-21K. IN1K-Self: self-supervised pre-training on ImageNet-1K with MoCo v3 [1]. The column labeled by † uses the same learning rate of 0.005 for the entire model, while the others use a learning rate of 0.01 for the classification layer.