How Well Does Self-Supervised Pre-Training Perform with Streaming ImageNet?

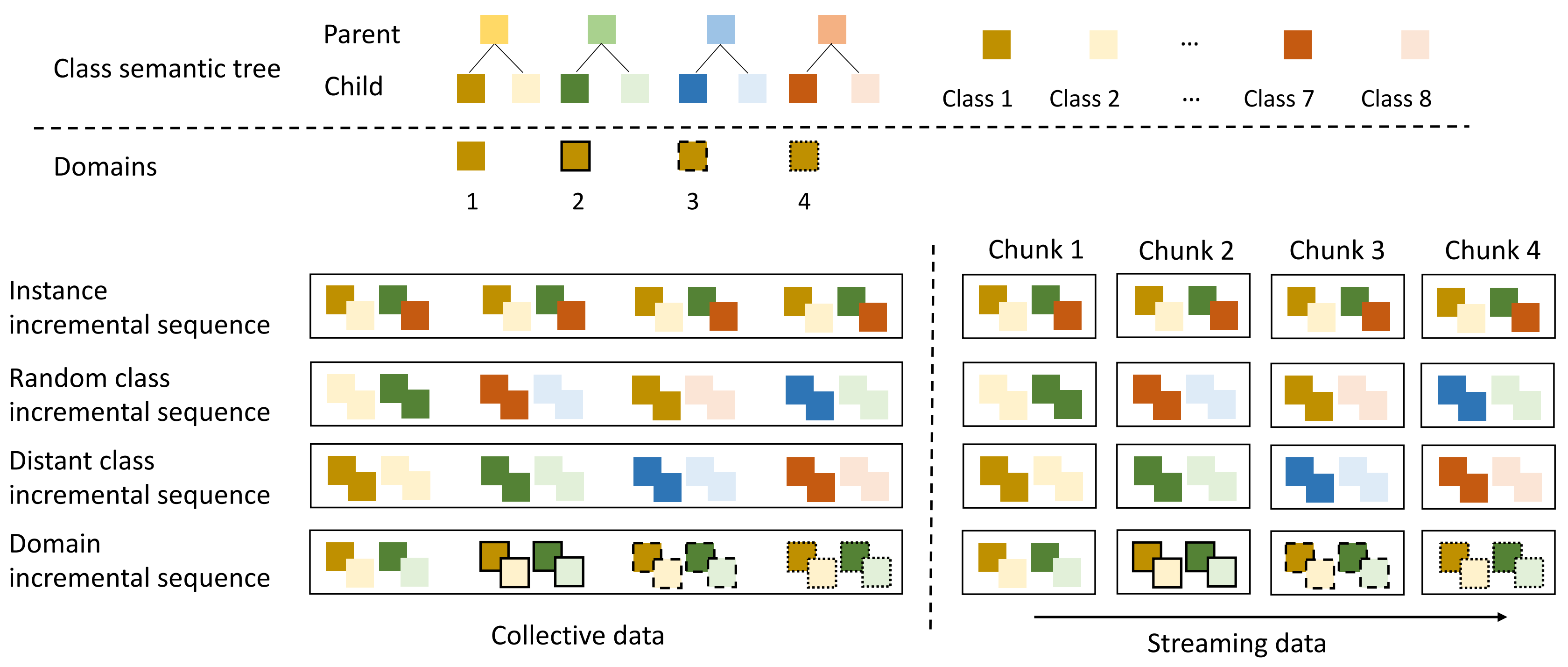
Dapeng Hu [1*], Shipeng Yan [2*], Qizhengqiu Lu [3], Lanqing Hong [4], Hailin Hu [3], Yifan Zhang [1], Zhenguo Li [4], Xinchao Wang [1], Jiashi Feng [1]

[1] National University of Singapore, [2] ShanghaiTech University

[3] AARC, Huawei Technologies, [4] Huawei Noah's Ark Lab

NEURAL INFORMATION PROCESSING SYSTEMS

1. Sequential Training vs. Joint Training



Class semantic tree

Parent

Child

Class 1 Class 2 ··· Class 7 Class 8

Domains

1 2 3 4

Instance incremental sequence

Random class incremental sequence

Distant class incremental sequence

Domain incremental sequence

Collective data

Chunk 1 Chunk 2 Chunk 3 Chunk 4

Streaming data

- In the upstream pre-training, we propose four types of streaming data with different degrees of distribution shift to mimic real-work data collection scenarios. MoCo-v2 is adopted as the pre-training method.
- In the downstream evaluation, we adopt three common downstream tasks, including many-shot evaluation and few-shot evaluation on 12 image classification tasks and object detection on Pascal VOC.

2. Dissection of Sequential Self-Supervised Pre-Training

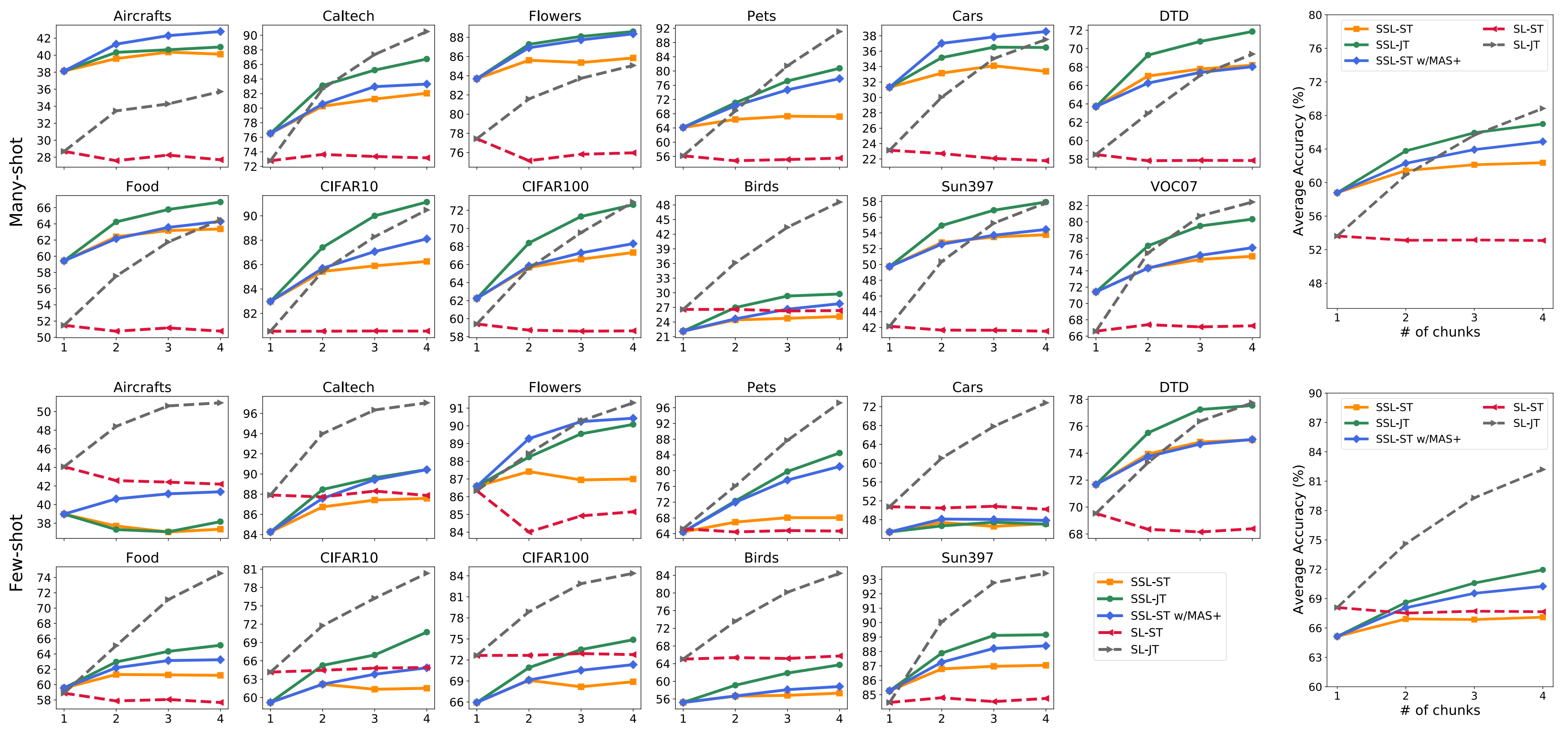


Table 1: Resource efficiency of considered SSL pre-training methods. We take the distant class incremental sequence as an example and report the training time (h) and required storage (GB) of the model pre-trained with each data chunk. Note that all the following statistics are recorded under the same hardware environment. The lower value means better efficiency.

Time (Storage) / Chunk	2	3	4
SSL-ST	**16.5 (35)**	**16.5 (35)**	**16.6 (35)**
SSL-ST W/Replay	17.0 (35)	18.5 (42)	20.0 (46)
SSL-ST w/MAS	18.2 (35)	18.1 (35)	18.1 (35)
SSL-ST w/MAS+	22.4 (39)	24.4 (42)	26.4 (46)
SSL-JT	31.1 (70)	46.5 (105)	66.6 (140)

Table 2: The comparison of pre-training methods in terms of the transfer performance gap between ST and JT models. We report the averaged accuracy gaps of linear evaluation across 12 downstream datasets. The lower, the better.

Accuracy gap (%) / Chunk	2	3	4
SL-ST (Instance)	2.26	3.27	4.83
SSL-ST (Instance)	**0.41**	**1.02**	**1.04**
SL-ST (Random)	5.63	8.73	10.68
SSL-ST (Random)	**0.42**	**0.94**	**1.13**
SL-ST (Distant)	7.77	12.50	15.75
SSL-ST (Distant)	2.34	3.81	4.62
SSL-ST w/MAS (Distant)	1.82	2.73	3.17
SSL-ST w/MAS+ (Distant)	**1.47**	**2.01**	**2.10**

- ST is more time-efficient and storage-saving then JT.
- Using continual learning methods introduces a tradeoff between efficiency and performance.

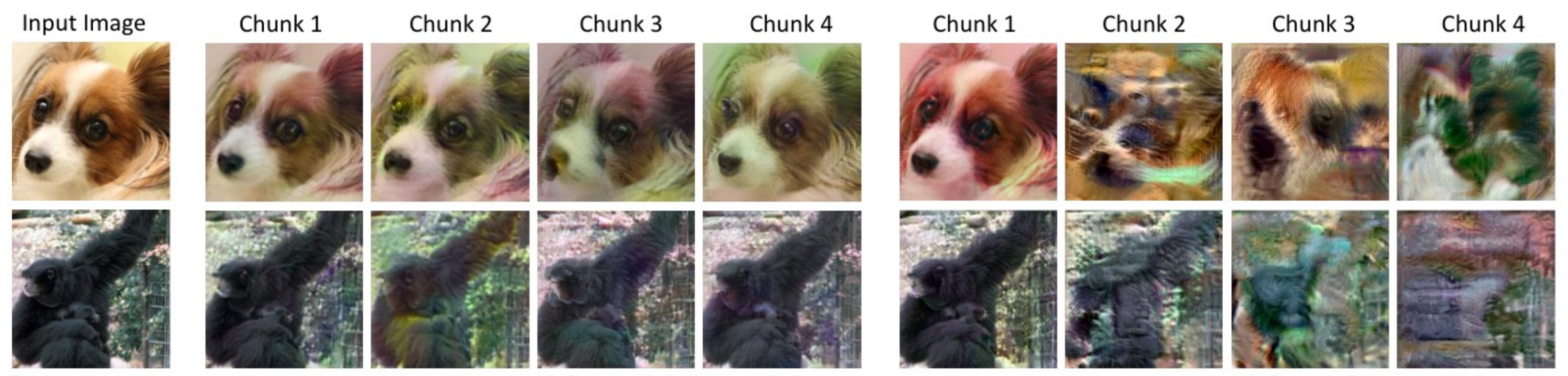
- For streaming data, severe distribution shift in streaming data challenges the sequential SSL.
- For downstream tasks, SSL representations are poor at few-shot transfer. Good upstream performances do not ensure good downstream transfer.
- For SSL methods, BYOL performs similarly to MoCo-v2 on the distant class incremental sequence.
- For continual learning methods, both replay and regularization are effective and can work together to improve the performance.

3. SSL Models Forget Less than SL Models

Backward transfer:

Data	Method	BWT(%)		FWT(%)	
		Top-1	Top-5	Top-1	Top-5
Instance	SL	-9.45	-5.46	**8.64**	2.81
	SSL	**3.61**	**3.60**	7.55	**8.63**
Random	SL	-20.63	-7.03	-0.34	0.01
	SSL	**-5.17**	**-1.36**	**11.05**	**4.52**
Distant	SL	-40.43	-28.66	4.90	0.47
	SSL	**-13.24**	**-11.06**	**11.01**	**3.66**

Feature reconstructions:



Raw image Images reconstructed from SSL models Images reconstructed from SL models

- Smaller negative transfer even positive transfer of SSL models.
- SSL representations retain more information in the sequential training process.

4. Our Hypothesis: SSL Models Have Flatter Minima

The sharpness metric based on representations:

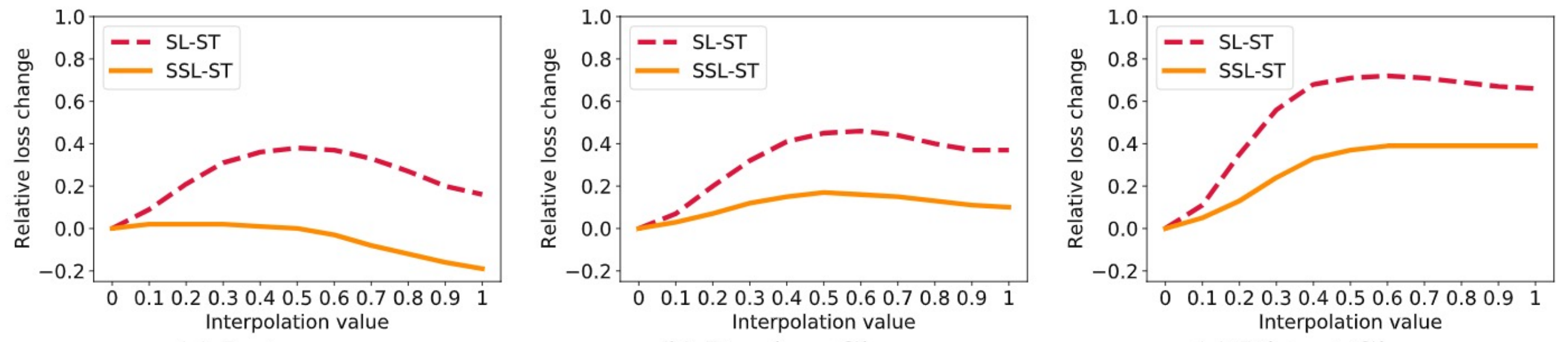
$$\mathcal{C}_\epsilon = \{\boldsymbol{z} \in \mathbb{R}^n : -\epsilon||\boldsymbol{\theta}||_2 \leq ||\boldsymbol{z}||_2 \leq \epsilon||\boldsymbol{\theta}||_2\}$$

$$\Phi_{\boldsymbol{\theta},f}(\epsilon) = \max_{\boldsymbol{\theta}' \in \mathcal{C}_\epsilon} g(\boldsymbol{\theta}') = \frac{f(\boldsymbol{\theta}) - \min_{\boldsymbol{\theta}'} f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})}$$

The sharpness results via perturbation:

	Instance		Random Class		Distant Class	
	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 0.1$	$\epsilon = 0.3$
SL	0.47	0.94	0.21	0.94	0.19	0.94
SSL	**0.14**	**0.68**	**0.08**	**0.66**	**0.06**	**0.71**

The sharpness results via interpolation:



(a) Instance (b) Random Class (c) Distant Class

Takeaways

- For streaming data with mild distribution shift, sequential SSL is a good choice. For streaming data with severe distribution shifts or longer sequences, sequential SSL with unsupervised parameter regularization and simple data replay is efficient.
- Compared with supervised learning (SL) models, SSL models consistently show smaller performance gaps between ST and JT. Our comprehensive investigation of learned representations demonstrates that sequential SSL models are less prone to catastrophic forgetting than SL models.
- Through the empirical analysis on the sharpness minima in the loss landscape, we find that SSL models have wider minima than SL models, which we argue is the probable reason for less forgetting of SSL models.

➢If you require any further information, feel free to contact me via: dapeng.hu@u.nus.edu