APPENDIX

A. Hyperparameter selection

To select hyperparameters, we employ a grid search strategy, using a randomly drawn 10% of the training data as the validation set. The considered hyperparameters are:

- Learning rate (η)
- Batch size (bsz)
- Number of start epochs (E_1)
- Number of epochs of t-th task ($E_{t\geq 2}$)
- Temperature for plasticity loss (τ): We use the same τ for both Focal Neural Collapse Contrastive (FNC^2) and Asymmetric SupCon loss [3]
- Focusing hyperparameters (γ) for FNC^2 loss
- Temperature for instance-wise relation distillation loss ($\mathcal{L}_{IRD}$): As in [3], we use different temperature hyperparameters for the past (κ_{past}) and current ($\kappa_{current}$) similarity vectors
- Temperature for sample-prototype relation distillation loss ($\mathcal{L}_{S-PRD}$): We utilize ζ_{past} for the past and $\zeta_{current}$ for the current similarity vectors
- Number of warm-up epochs in hardness-softness distillation loss ($\mathcal{L}_{HSD}$) (e_0)

The corresponding search space of these hyperparameters are provided in Tab. 1. The selections of these hyperparameters are based on the average test accuracy over five independent trials, and the final chosen values are detailed in Tab. 2. For the sake of conciseness and to maintain focus, we omit those hyperparameters previously discovered in the literature.

Focusing hyperparameter (γ). In the FNC^2 loss function, γ plays a crucial role in determining the level of focus on hard samples (i.e., positive samples that are far from the anchor or their prototypes). To explore this role and select the most suitable γ for each dataset, we conduct experiments across different datasets to observe how the the performance of our method changes as γ varies. The test accuracy results in Fig. 1 show that our method performs best at different values of γ for each dataset. Specifically, as reported in Tab. 2, the chosen γ for the Seq-Cifar-100 and Seq-Tiny-Imagenet datasets ($\gamma = 4$ for both) are larger than that for the Seq-Cifar-10 dataset ($\gamma = 1$). This difference arises because Seq-Cifar-100 and Seq-Tiny-ImageNet have a large number of classes per task (both have 20 classes/task), which increases the likelihood of samples being close to the prototypes of other class clusters. In contrast, the Seq-Cifar-10 dataset has only 2 classes each task, making it less complex and not requiring a large γ.

Hyperparameter	Values
η	{0.1, 0.5, 1.0}
bsz	{256, 512}
E_1	{500}
$E_{t\geq 2}$	{50, 100}
τ	{0.1, 0.5, 1.0}
γ	{0, 1, 2, 4, 7, 10}
κ_{past}	{0.01, 0.05, 0.1}
$\kappa_{current}$	{0.1, 0.2}
ζ_{past}	{0.01, 0.05, 0.1}
$\zeta_{current}$	{0.1, 0.2}
e_0	{10, 20, 30}

Table 1. Search spaces of hyperparameters.

Method	Buffer size	Dataset	Hyperparameter
Our	0, 200, 500	Seq-Cifar-10	η: 0.5, γ: 1, bsz: 512, E_1: 500, $E_{t\geq 2}$: 100, τ: 0.5, e_0: 30, κ_{past}: 0.01, $\kappa_{current}$: 0.2, ζ_{past}: 0.01, $\zeta_{current}$: 0.2
	0, 200, 500	Seq-Cifar-100	η: 0.5, γ: 4, bsz: 512, E_1: 500, $E_{t\geq 2}$: 100, τ: 0.5, e_0: 30, κ_{past}: 0.01, $\kappa_{current}$: 0.2, ζ_{past}: 0.1, $\zeta_{current}$: 0.2
	0, 200, 500	Seq-Tiny-ImageNet	η: 0.1, γ: 4, bsz: 512, E_1: 500, $E_{t\geq 2}$: 50, τ: 0.5, e_0: 20, κ_{past}: 0.1, $\kappa_{current}$: 0.1, ζ_{past}: 0.1, $\zeta_{current}$: 0.2
Co^{2}L	0, 200, 500	Seq-Cifar-100	η : 0.5, bsz: 512, E_1: 500, $E_{t\geq 2}$: 100, τ: 0.5, κ_{past}: 0.01, $\kappa_{current}$: 0.2
	0	Seq-Tiny-ImageNet	η : 0.1, bsz: 512, E_1: 500, $E_{t\geq 2}$: 50, τ: 0.5, κ_{past}: 0.1, $\kappa_{current}$: 0.1

Table 2. Selected hyperparameters in our experiments.

B. Additional Experiments

B.1. Average accuracy results with buffer size 500

In addition to the results with small buffer sizes (0 and 200), we run experiments with a buffer size of 500 across different datasets to further assess the effectiveness of our method with a larger buffer. As shown in Tab. 3, although our method does not surpass state-of-the-art methods, it achieves results close to them on Seq-Cifar-10 and Seq-Tiny-ImageNet, underperforming only on Seq-Cifar-100 compared to GCR [8]. This further demonstrates that our method, aside from excelling in memory-free and small buffer settings, remains effective with larger buffers.

B.2. Average forgetting results

We utilize the Average Forgetting metric as defined in [4] to quantify how much information the model has forgotten about previous tasks, which as

$$F = \frac{1}{T-1}\sum_{i=1}^{T-1} max_{t\in\{1,\dots,T-1\}}(A_{t,i} - A_{T,i}) \qquad (1)$$

Tab. 4 report the average forgetting results of our method compared to all other baselines. The results show that our method can effectively mitigate forgetting, especially even without using additional buffers.

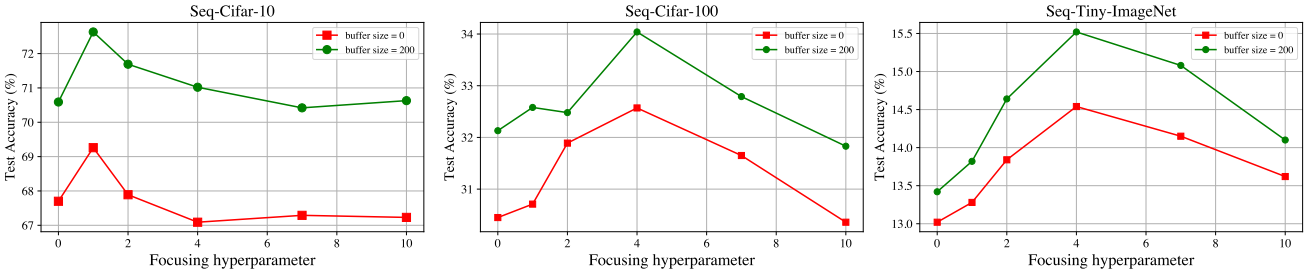

Figure 1. Test accuracy over different values of γ.

Buffer	Dataset	Seq-Cifar-10		Seq-Cifar-100		Seq-Tiny-ImageNet	
	Scenario	Class-IL	Task-IL	Class-IL	Task-IL	Class-IL	Task-IL
500	ER [7]	57.74±0.27	93.61±0.27	27.66±0.61	66.23±1.52	9.99±0.29	48.64±0.46
	iCaRL [6]	47.55±3.95	88.22±2.62	33.25±1.25	58.16±1.76	9.38±1.53	31.55±3.27
	GEM [5]	26.20±1.26	92.16±0.64	25.54±0.65	66.31±0.86	-	-
	GSS [1]	49.73±4.78	91.02±1.57	21.92±0.34	60.28±1.18	-	-
	DER [2]	70.51±1.67	93.40±0.39	41.36±1.76	71.73±0.74	17.75±1.14	51.78±0.88
	Co^{2}L [3]	74.26±0.77	95.90±0.26	37.02±0.76	62.44±0.36	20.12±0.42	53.04±0.69
	GCR [8]	74.69±0.85	94.44±0.32	**45.91±1.30**	**71.64±2.10**	19.66±0.68	52.99±0.89
	CILA [9]	**76.03±0.79**	**96.40±0.21**	-	-	**20.64±0.59**	**54.13±0.72**
	Ours	75.51±0.52	96.14±0.25	40.25±0.58	65.85±0.44	20.31±0.34	53.46±0.59

Table 3. Additional results with buffer size 500 (best results in each column are bold).

Buffer	Dataset	Seq-Cifar-10		Seq-Cifar-100		Seq-Tiny-ImageNet	
	Scenario	Class-IL	Task-IL	Class-IL	Task-IL	Class-IL	Task-IL
0	Co^{2}L [3]	35.81±1.08	14.33±0.87	66.51±0.28	39.63±0.62	62.80±0.77	39.54±1.08
	Ours	**23.85±0.30**	**4.72±0.28**	**52.03±0.63**	**36.20±0.48**	**53.97±0.63**	**37.57±0.88**
200	ER [7]	59.30±2.48	6.07±1.09	75.06±0.63	27.38±1.46	76.53±0.51	40.47±1.54
	GEM [5]	80.36±5.25	9.57±2.05	77.40±1.09	29.59±1.66	-	-
	GSS [1]	72.48±4.45	8.49±2.05	77.62±0.76	32.81±1.75	-	-
	iCARL [6]	**23.52±1.27**	25.34±1.64	**47.20±1.23**	36.20±1.85	**31.06±1.91**	42.47±2.47
	DER [2]	35.79±2.59	6.08±0.70	62.72±2.69	25.98±1.55	64.83±1.48	40.43±1.05
	Co^{2}L [3]	36.35±1.16	6.71±0.35	67.82±0.41	38.22±0.34	73.25±0.21	47.11±1.04
	GCR [8]	32.75±2.67	7.38±1.02	57.65±2.48	**24.12±1.17**	65.29±1.73	40.36±1.08
	CILA [9]	-	-	-	-	-	-
	Ours	25.24±0.69	**4.28±0.32**	52.40±0.83	33.66±0.24	52.07±0.46	**33.76±0.58**
500	ER [7]	43.22±2.10	3.50±0.53	67.96±0.78	17.37±1.06	75.21±0.54	30.73±0.62
	GEM [5]	78.93±6.53	5.60±0.96	71.34±0.78	20.44±1.13	-	-
	GSS [1]	59.18±4.00	6.37±1.55	74.12±0.42	26.57±1.34	-	-
	iCARL [6]	28.20±2.41	22.61±3.97	40.99±1.02	27.90±1.37	**37.30±1.42**	39.44±0.84
	DER [2]	24.02±1.63	3.72±0.55	49.07±2.54	25.98±1.55	59.95±2.31	28.21±0.97
	Co^{2}L [3]	25.33±0.99	3.41±0.80	51.23±0.65	26.30±0.57	65.15±0.26	39.22±0.69
	GCR [8]	**19.27±1.48**	**3.14±0.36**	**39.20±2.84**	**15.07±1.88**	56.40±1.08	27.88±1.19
	CILA [9]	-	-	-	-	-	-
	Ours	22.59±1.02	3.21±0.25	41.66±0.78	24.84±0.91	46.08±0.56	**26.45±0.79**

Table 4. Average forgetting (lower is better) across five independent trials: Comparison of our method with all baselines in continual learning.

References

[1] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *NeurIPS*, 2019. 2

[2] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*, 2020. 2

[3] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co^{2}L: Contrastive Continual Learning. In *ICCV*, 2021. 1, 2

[4] Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, 2018. 1

[5] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NeurIPS*, 2017. 2

[6] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 2

[7] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *ICLR*, 2019. 2

[8] Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *CVPR*, 2022. 1, 2

[9] Yichen Wen, Zhiquan Tan, Kaipeng Zheng, Chuanlong Xie, and Weiran Huang. Provable contrastive continual learning. In *ICML*, 2024. 2