






# Faster convergence and Uncorrelated gradients in Self-Supervised Online Continual Learning

Koyo Imai<sup>1</sup>, Naoto Hayashi<sup>1</sup>, Tsubasa Hirakawa<sup>1</sup>, Takayoshi Yamashita<sup>1</sup>,  
and Hironobu Fujiyoshi<sup>1</sup>

Chubu University, 1200 Matsumoto-cho, Kasugai, Aichi, Japan  
{kouyou, hayashi29, hirakawa}@mprg.cs.chubu.ac.jp,  
{takayoshi, fujiyoshi}@isc.chubu.ac.jp

## 1 Data stream

In this section, we describe the details of the data streams used in the evaluation experiments. The datasets used were CIFAR-10, CIFAR-100 [5], ImageNet-100 [2], and CORe50 [6], and the data stream was created following [10]. A data stream is created by dividing the dataset into separate subsets for each task and connecting them. Note that all scenarios used in the evaluation experiments are class-incremental scenarios [9], where tasks represent different data distributions. The classes included in each task are different, and the data distribution changes as the task progresses. The following subsections describe the Sequential, Sequential blurred boundaries, and Sequential imbalance data streams that were created.

### 1.1 Data stream setting

**Sequential data stream.** The Sequential data stream (Seq) is a data stream with progressively increasing classes. The settings for each dataset are shown in Table 4. In each dataset, the number of data per class is uniform, and the number of data included per task is the same.

Table 1: Sequential stream

	Number of data		Number of classes		Number of Tasks
	/ class	/ task	/ task	/ stream	
CIFAR-10	5,000	5,000	1	10	10
CIFAR-100	500	2,500	5	100	20
ImageNet-100	1,200	12,000	10	100	10
CORe50	3,000	3,000	1	50	50

Table 2: Sequential blurred boundaries stream

	Number of data		Number of classes		Number of tasks	Mixing Ratio of Task Boundaries
	/class	/task	/task	/stream		
CIFAR-10	5,000	5,000	1	10	10	0.25
CIFAR-100	500	2,500	5	100	20	0.25
ImageNet-100	1,200	12,000	10	100	10	0.25

Table 3: Sequential imbalance stream

	Number of data		Number of classes		Number of tasks
	/class	/task	/task	/stream	
CIFAR-10	5,000 or 2,500	5,000 or 2,500	1	10	10
CIFAR-100	0 ~ 500	2,500 or 1,200	5	100	20
ImageNet-100	0 ~ 1,200	12,000 or 6,000	10	100	10

**Sequential blurred boundaries data stream.** The Sequential Blurred Boundaries data stream (Seq-bl) is a data stream with ambiguous task boundaries. The settings for each dataset are shown in Table 2. The Mixing Ratio of Task Boundaries is the ratio of mixing the data of task  $t$  with the data of task  $t + 1$ . Settings such as the number of data and classes are the same as for the Sequential data stream. The difference from the Sequential data stream is that the Mixing Ratio of Task Boundaries is set to 0.25. Seq-bl blurs task boundaries by mixing the last 25% data in task  $t$  with the first 25% data in task  $t + 1$ .

**Sequential imbalance data stream.** The Sequential Imbalance data stream (Seq-im) is a data stream that contains an unbalanced number of data per task. The settings for each dataset are shown in Table 3. The number of data varies from task to task, making the change in data distribution irregular.

## 2 Implementation Details

### 2.1 Implementation Details of conventional method

We describe the establishment of baselines used in the evaluation experiments. For these experiments, we use the conventional methods MinRed [7], SCALE [10], and EMP-SSL [8] as baselines. Each method uses ResNet-18 [4] with 512 output dimensions as the network. The hyperparameters used for training each method are shown in Table 4. Hyperparameters and data augmentations not

Table 4: Hyperparameters of baseline

Parameter	Explain	Value
$lr$	Learning rate (MinRed, SCALE)	0.03
	Learning rate (EMP-SSL)	0.01
$b$	Batch size for data stream and replay buffer	100
	Rehearsal iterations (EMP-SSL)	3
$K$	Rehearsal iterations (SCALE)	20
	Rehearsal iterations (MinRed)	40
$ \mathcal{M} $	Buuffer size	1,024

shown in Table 4 follow the implementation details of each paper. In the evaluation experiments, the conventional methods are trained with the above settings unless otherwise noted.

## 2.2 Implementation Details of proposed method

The setup of the proposed method used in the evaluation experiment is described as follows. The proposed method uses ResNet-18 with 512 output dimensions as the network. A projector with a hidden layer dimension of 4,096 and an output layer of 1,024 is connected behind ResNet-18. The proposed method uses LARS as the optimization method, setting  $\eta$  to 0.005 and weight decay to  $1e-4$ .

**Hyperparameter.** The basic hyperparameter settings used for training the proposed method are shown in Table 5. The hyperparameters of the proposed method in the evaluation experiments follow Table 5 unless otherwise noted.

**Data Augmentation.** The data augmentations used in the proposed method are the same as those used in VICReg [1] and BYOL [3]. The following operations are performed sequentially to generate  $N$  views.

- Random cropping with an area uniformly sampled with a size ratio of 0.25. The cropped images are resized to  $32 \times 32$  for CIFAR,  $128 \times 128$  for CORE50, and  $224 \times 224$  for ImageNet-100.
- Random horizontal flip with a probability of 0.5.
- Color jitter with brightness 0.4, contrast 0.4, saturation 0.4, and hue 0.2, applied with a probability of 0.8.
- Grayscale with a probability of 0.2.
- Gaussian blur with a probability of 0.1.
- Solarization with a probability of 0.1.

Table 5: Hyperparameters of proposed method

Parameter	Explain	Value
$lr$	Learning rate	0.01
$N$	Number of crops	20
$b$	Batch size for data stream and replay buffer	100
$\lambda$	Weights for Multi-Crop Contrast loss	200
$\tau$	Temperature for Multi-Crop Contrastive loss	0.07
$K$	Rehearsal iterations	3
$ \mathcal{M} $	Buuffer size	1,024
$\alpha$	Moving average coefficient	0.5

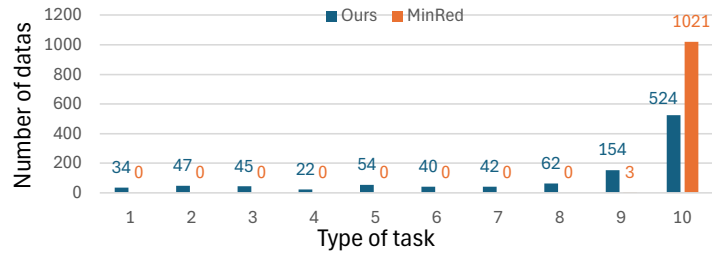


Fig. 1: Number of data in buffer at the end of Seq-CIFAR10 training when only the data selection method is changed. The buffer size is 1,024.

### 3 Ablation Study

#### 3.1 Data Selection

We investigate the effectiveness of the proposed method of selecting data to be stored in the replay buffer using average features. Figure. 1 shows the comparison results between using Cosine similarity of the average feature and MinRed as the data selection method. From Figure. 1, we see that the proposed data selection method keeps a greater variety of data in the buffer. Therefore, we believe that retaining knowledge learned in the past using data in the buffer is easier than with the MinRed data selection.

#### 3.2 TCR Loss

An accuracy comparison on Seq-ImageNet-100 without TCR Loss is shown in Table 6. The effectiveness of introducing TCR Loss is confirmed by the fact that removing TCR Loss shows a decrease in accuracy, as shown in Table 6.

Table 6: TCR Loss ablation study

	w/ TCR Loss	w/o TCR Loss
Ours	25.81	23.19

## References

1. Bardes, A., Ponce, J., LeCun, Y.: VICReg: Variance-invariance-covariance regularization for self-supervised learning. In: International Conference on Learning Representations (2022), <https://openreview.net/forum?id=xm6YD62D1Ub>
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
3. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Dorsch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems* **33**, 21271–21284 (2020)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
5. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
6. Lomonaco, V., Maltoni, D.: Core50: a new dataset and benchmark for continuous object recognition. In: Levine, S., Vanhoucke, V., Goldberg, K. (eds.) Proceedings of the 1st Annual Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 78, pp. 17–26. PMLR (13–15 Nov 2017), <https://proceedings.mlr.press/v78/lomonaco17a.html>
7. Purushwalkam, S., Morgado, P., Gupta, A.: The challenges of continuous self-supervised learning. In: European Conference on Computer Vision. pp. 702–721. Springer (2022)
8. Tong, S., Chen, Y., Ma, Y., Lecun, Y.: Emp-ssl: Towards self-supervised learning in one training epoch. arXiv preprint arXiv:2304.03977 (2023)
9. Van de Ven, G.M., Tolias, A.S.: Three scenarios for continual learning. arXiv preprint arXiv:1904.07734 (2019)
10. Yu, X., Guo, Y., Gao, S., Rosing, T.: Scale: Online self-supervised lifelong learning without prior knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 2484–2495 (June 2023)