ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448032

UNSUPERVISED CONTINUAL LEARNING OF IMAGE REPRESENTATION VIA REMEMORY-BASED SIMSIAM

Feifei Fu*, *Yizhao Gao**, *Zhiwu Lu*$^{\star *}$, *Haoran Wu*†, *Shiqi Zhao*†

* Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
† China Unicom Research Institute, Beijing, China

ABSTRACT

Unsupervised continual learning (UCL) of image representation has garnered attention due to practical need. However, recent UCL methods focus on mitigating the catastrophic forgetting with a replay buffer (i.e., rehearsal-based strategy), which needs much extra storage. To overcome this drawback, we propose a novel rememory-based SimSiam (RM-SimSiam) method to reduce the dependency on replay buffer. The core idea of RM-SimSiam is to store and remember the old knowledge with a data-free historical module instead of replay buffer. Specifically, this historical module is designed to store the historical average model of all previous models (the memory process) and then transfer the knowledge of the historical average model to the new model (the rememory process). To further improve the rememory ability of RM-SimSiam, we devise an enhanced SimSiam-based contrastive loss by aligning the representations outputted by the historical and new models. Extensive experiments on three benchmarks demonstrate the effectiveness of our RM-SimSiam.

Index Terms— Representation learning, Unsupervised continual learning, Rememory, Catastrophic forgetting

1. INTRODUCTION

Continual learning [1] can be divided into two categories according to whether the training data is labeled or not: supervised continual learning (SCL), and unsupervised continual learning (UCL). SCL has been studied extensively in the past few years [2, 3, 4]. However, motivated by the practical need in real-world application scenarios, researchers have started to turn their attention to the unsupervised field: representation learning with unlabeled image data on sequential tasks (i.e., UCL). Recent UCL methods [5, 6, 7] have achieved promising results by exploring various unsupervised strategies to mitigate the catastrophic forgetting [8]. However, most of them focus on utilizing a large replay buffer to store previous data (i.e., rehearsal-based strategy), which needs much extra storage and thus limits their practical applications. For example, [5] mitigates the forgetting by storing a constant number of images per class and requiring an extra sample queue to store negative samples of old data, which is extremely storage-wasting. In [6], the two better techniques DER [3] and LUMP also depend on a replay buffer to mitigate the forgetting by storing old data.

To overcome the drawback of most recent UCL methods, based on unsupervised contrastive learning via SimSiam [9], we propose a novel rememory-based SimSiam (RM-SimSiam) method to reduce the dependency on replay buffer. Analogous to the memory mechanism of human brain [10, 11], the core idea of RM-SimSiam is to store and remember the old knowledge with a data-free historical module (instead of replay buffer that stores old data directly). Specifically, our RM-SimSiam model mainly consists of two modules: hist-module (i.e., historical module) and new-module (see Figure 1). The hist-module is designed to store the historical average model of all previous models (i.e., the memory process) and then transfer the knowledge of the historical average model to the new-module (i.e., the rememory process) for retaining the previously learned knowledge. By such memory & rememory process, the old knowledge can be effectively consolidated (memorized) and remembered (rememorized) throughout the optimization trajectory, thus ensuring that RM-SimSiam can mitigate the forgetting of the old knowledge when learning a new task.

Furthermore, to improve the rememory ability of RM-SimSiam, we devise an enhanced SimSiam-based contrastive loss by aligning the feature representations outputted by the historical and new models. Such alignment mechanism is different from that in the latest work [7] on UCL. [7] exploits the distillation mechanism to align the representations of the current and past states by saving the model checkpoint of the past state. In contrast, we align the representations of the historical average model (of all previous models) and new model in each iteration process. Note that the largest difference between [7] and our RM-SimSiam still lies in that the novel rememory process is included in RM-SimSiam to learn a new task well while mitigating forgetting, but such rememory process is ignored in [7].

Our main contributions are three-fold: **(1)** We propose a novel rememory-based method termed RM-SimSiam for unsupervised continual learning of image representation by storing and remembering the old knowledge with a data-free historical module to reduce the dependency on replay buffer. **(2)**

*Corresponding author

To effectively rememory the knowledge of previous tasks, we design a hist-module by storing the knowledge of previous models and transferring the knowledge of previous models to the new model. To further improve the rememory ability of our RM-SimSiam, we devise an enhanced SimSiam-based contrastive loss by aligning the representations outputted by the historical and new models. **(3)** Extensive experiments on three benchmarks show that our RM-SimSiam achieves new state-of-the-art under the UCL setting.

2. METHODOLOGY

2.1. Problem Definition

Given a sequence of tasks $T = \{T_1, T_2, ..., T_n\}$, where n denotes the number of tasks. Each task T_t $(1 \leq t \leq n)$ from T has a task-specific training set $D_t = \{x_i, y_i\}_{i=1}^{N_t}$, where x_i denotes an image, y_i denotes the ground-truth class label of x_i, and N_t denotes the number of training samples. Given that D_t is drawn from the i.i.d. distribution $P_t(x, y)$, we assume that any pair of tasks T_t and T_{t+j} $(1 \leq j \leq n - t)$ have different distributions: $P_t(x, y) \neq P_{t+j}(x, y)$. In addition, for each T_t, its validation and test sets can be defined similarly.

Since UCL is considered (but not SCL) in this paper, there is no labeled samples during training. That is, for each task T_t, it has an unlabeled training set $U_t = \{x_i\}_{i=1}^{N_t}$ with N_t training samples (but its validation and test sets have labeled samples). The learning process for UCL is thus given as follows: (1) The feature representations of the training samples are learned on the set of sequential tasks; (2) K-nearest neighbor (KNN) classifier [12] is performed on the validation set to obtain the classification accuracy for hyperparameter tuning; (3) The performance on the test set is evaluated based on KNN classifier, following the setup in [13, 9].

2.2. SimSiam

SimSiam [9] is a simple yet effective method for unsupervised representation learning, which mainly includes an encoder f and a predictor head h, just like the new-module in Figure 1. The encoder f consists of the backbone ResNet18 [14] (without pretraining), and the predictor head h consists of multilayer perceptron (MLP) layers. Given an input image x, the output of the encoder f is $z \triangleq f(x)$, and the output of the predictor h is $p \triangleq h(z) \triangleq h(f(x))$. For the two augmented views x_1 and x_2 of the input image x, SimSiam chooses to learn the feature representations by minimizing the cosine-distance between the output of one view's predictor (e.g., $x_1 \longrightarrow p_1 \triangleq h(f(x_1))$) and the output of the other view's encoder (e.g., $x_2 \longrightarrow z_2 \triangleq f(x_2)$) and vice versa.

According to [9], a symmetric contrastive loss L_{sim} is employed to learn accurate representations, it is defined as:

$$L_{sim} = \frac{1}{2}D(p_1, z_2) + \frac{1}{2}D(p_2, z_1), \tag{1}$$

$$D(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}, \tag{2}$$

where D is a cosine-distance function, and $\|\cdot\|_2$ is l_2-norm. Since a stop-gradient operation $\mathrm{sg}(\cdot)$ is imposed on z to prevent model collapse, L_{sim} is reformulated as:

$$L_{sim} = \frac{1}{2}D(p_1, \mathrm{sg}(z_2)) + \frac{1}{2}D(p_2, \mathrm{sg}(z_1)). \tag{3}$$

When SimSiam is applied to continual learning, given an input image $x_{i,t}$ from the task T_t, the symmetric contrastive loss L_{sim} is defined as:

$$L_{sim} = \frac{1}{2}D(p_{i,t}^1, \mathrm{sg}(z_{i,t}^2)) + \frac{1}{2}D(p_{i,t}^2, \mathrm{sg}(z_{i,t}^1)), \tag{4}$$

where the two augmented views of $x_{i,t}$ are $x_{i,t}^1$ and $x_{i,t}^2$, the encoder output $z_{i,t}^j \triangleq f(x_{i,t}^j)$ $(j = 1, 2)$, and the predictor output $p_{i,t}^j \triangleq h(f(x_{i,t}^j))$ $(j = 1, 2)$.

2.3. RM-SimSiam

Inspired by the memory mechanism of human brain and based on unsupervised contrastive learning via SimSiam model, we propose the novel rememory-based SimSiam (RM-SimSiam) on UCL. As illustrated in Figure 1, RM-SimSiam mainly has two modules: new-module and hist-module (historical module). Among them, the new-module is mainly used to learn the knowledge of the current new task (e.g., T_t), the hist-module is mainly used to retain the learned knowledge in the previous tasks (e.g., $T_1, T_2, ..., T_{t-1}$). With the proposed rememory mechanism and the enhanced SimSiam-based contrastive loss, our RM-SimSiam can learn new knowledge well while mitigating the catastrophic forgetting.

Rememory Mechanism for UCL. To mitigate the catastrophic forgetting problem under the UCL setting, we propose the rememory mechanism to consolidate (memory) and remember (rememory) the previously learned knowledge. Specifically, we design the hist-module to retain the old knowledge by storing the historical average model of all previous models (i.e., the memory process) and then transferring the knowledge of the historical average model to the new-module (i.e., the rememory process). As shown in Figure 1, in the new-module and hist-module, the encoders are respectively denoted as f (with parameters θ_f^e) and q (with parameters θ_q^e), and the predictor heads respectively as h (with parameters θ_h^p) and g (with parameters θ_g^p). To consolidate the learned knowledge of previous tasks, we update the parameters θ_q^e, θ_g^p of the hist-module by transferring the parameters θ_f^e, θ_h^p of the new-module, which is called the memory process. In turn, to remember the previously learned knowledge, we transfer the parameters of the hist-module to the new-module, which is called the rememory process. These two transfer processes constitute our rememory mechanism. Given the transfer coefficient m, the two transfer processes are uniformly defined as:

$$\theta_i^e = m \cdot \theta_i^e + (1 - m) \cdot \theta_j^e, \;\; i, j \in \{f, q\}, \; i \neq j, \tag{5}$$

$$\theta_i^p = m \cdot \theta_i^p + (1 - m) \cdot \theta_j^p, \;\; i, j \in \{h, g\}, \; i \neq j, \tag{6}$$

where the parameters θ_q^e, θ_g^p of the hist-module have no gradient back-propagation.

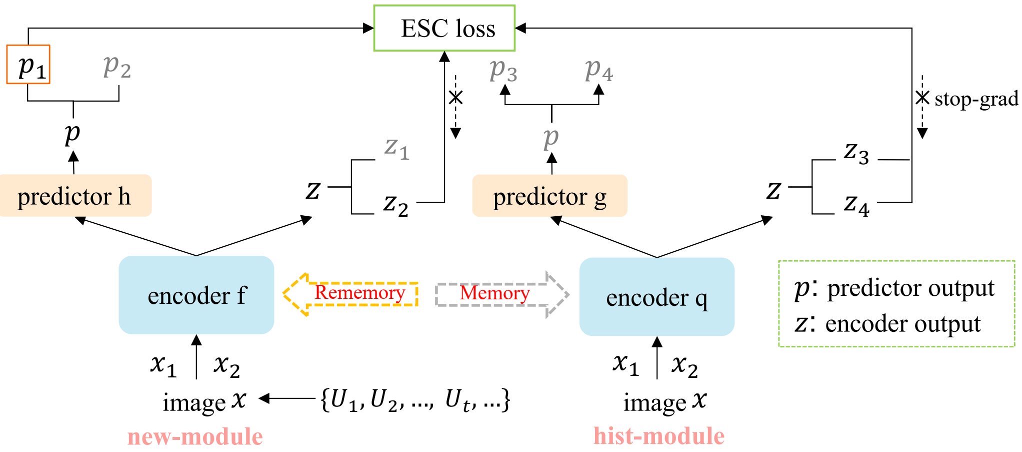

Fig. 1. Overview of our RM-SimSiam, which mainly consists of new-module and hist-module. The rememory mechanism is applied between them to learn the new knowledge well while retaining the old knowledge. The enhanced SimSiam-based contrastive (ESC) loss for model optimization is defined by taking both the historical and new models into consideration.

Enhanced SimSiam-based Contrastive Loss. Further, to improve the rememory ability of RM-SimSiam, we propose an enhanced SimSiam-based contrastive (ESC) loss by aligning the feature representations outputted by the historical and new models. Concretely, given an input image $x_{i,t}$, the new-module and hist-module take two randomly-augmented views $x_{i,t}^1, x_{i,t}^2$ of $x_{i,t}$ as inputs, and produce the corresponding encoder outputs $\{z_{i,t}^j\}$ and predictor outputs $\{p_{i,t}^j\}$ ($j = 1, 2$ for the new-module and $j = 3, 4$ for the hist-module), as shown in Figure 1. To better retain the previously learned knowledge, we add a new SimSiam-style contrastive loss L_{hist} on top of the original SimSiam loss L_{sim} given by Eq. (4). Formally, by taking the outputs of the two views in the hist-module as guidance, we can define L_{hist} (with a similar form to L_{sim}) as follows:

$$\begin{aligned} L_{hist} = &\frac{1}{2}D(p_{i,t}^1, z_{i,t}^3) + \frac{1}{2}D(p_{i,t}^3, z_{i,t}^1) + \frac{1}{2}D(p_{i,t}^1, z_{i,t}^4) \\ &+ \frac{1}{2}D(p_{i,t}^4, z_{i,t}^1) + \frac{1}{2}D(p_{i,t}^2, z_{i,t}^3) + \frac{1}{2}D(p_{i,t}^3, z_{i,t}^2) \\ &+ \frac{1}{2}D(p_{i,t}^2, z_{i,t}^4) + \frac{1}{2}D(p_{i,t}^4, z_{i,t}^2) + \frac{1}{2}D(p_{i,t}^3, z_{i,t}^4) \\ &+ \frac{1}{2}D(p_{i,t}^4, z_{i,t}^3). \end{aligned} \tag{7}$$

Noticing the non-gradient property of the hist-module, we further impose the stop-gradient operation $\mathrm{sg}(\cdot)$ on z. In this way, we can simplify the above contrastive loss L_{hist} as:

$$\begin{aligned} L_{hist} \triangleq &\frac{1}{2}D(p_{i,t}^1, \mathrm{sg}(z_{i,t}^3)) + \frac{1}{2}D(p_{i,t}^2, \mathrm{sg}(z_{i,t}^3)) \\ &+ \frac{1}{2}D(p_{i,t}^1, \mathrm{sg}(z_{i,t}^4)) + \frac{1}{2}D(p_{i,t}^2, \mathrm{sg}(z_{i,t}^4)). \end{aligned} \tag{8}$$

By combining L_{sim} and L_{hist}, our enhanced SimSiam-based contrastive (ESC) loss is defined as:

$$L_{esc} = L_{sim} + \gamma L_{hist}, \tag{9}$$

where γ is the weight hyperparameter, L_{sim} is the original SimSiam loss (see Eq. 4).

3. EXPERIMENTS

3.1. Experimental Setup

Datasets. Three classical datasets are selected for performance evaluation: **(1) SPLIT CIFAR-10** (S-CIFAR-10) [15] is split into 5 tasks (2 classes per task). Each class has 6,000 color images of 32 ∗ 32, of which 5,000 are used for training and 1,000 for testing. **(2) SPLIT CIFAR-100** (S-CIFAR-100) [15] is split into 20 tasks (5 classes per task). Each class has 600 color images of 32 ∗ 32, of which 500 are used for training and 100 for testing. **(3) SPLIT Tiny-IMAGENET** (S-Tiny-IMAGENET) [16] is a subset of ImageNet [17], the first 100 classes of which are used. Each class has 500 color images (the image size is 64 ∗ 64) for training and 50 images for testing. The task split is the same as that of S-CIFAR-100.

Implementation Details. Our RM-SimSiam adopts the Stochastic Gradient Descent (SGD) optimizer, with the learning rate 0.03 for S-CIFAR-10/S-CIFAR-100 and 0.035 for S-Tiny-IMAGENET. We set the batch size to 128. We set $m = 0.99$ and $\gamma = 1$. To explore the complementarity between the rehearsal-based method and our RM-SimSiam, we combine our RM-SimSiam with the Mixup strategy [18]. Following [6], the two metrics average accuracy (acc) and average forgetting (fg) over three independent runs are reported for performance evaluation. The code is available at link .

3.2. Main Results

We compare our proposed RM-SimSiam against other state-of-the-art methods under the UCL setting on the three benchmark datasets, as shown in Table 1. For fair comparison, both RM-SimSiam with memory buffer (denoted as RM-SimSiam) and RM-SimSiam without memory buffer (denoted as RM-SimSiam*) are considered. From Table 1, it can be observed that: **(1)** Our RM-SimSiam without memory buffer (i.e., RM-SimSiam*) leads to better results than most of the other UCL methods on all three benchmark datasets, demonstrating the effectiveness of our proposed RM-SimSiam. **(2)** When the memory buffer is used exactly

Table 1. Comparison to the state-of-the-arts under the UCL setting in terms of acc and fg over three independent runs. All UCL methods (with the same backbone ResNet18) are trained from scratch. * denotes our RM-SimSiam without buffer.

Method	S-CIFAR-10		S-CIFAR-100		S-Tiny-IMAGENET	
	acc (↑)	fg (↓)	acc (↑)	fg (↓)	acc (↑)	fg (↓)
FINETUNE	90.11 (±0.12)	5.42 (±0.08)	75.42 (±0.78)	10.19 (±0.37)	71.07 (±0.20)	9.48 (±0.56)
PNN [19]	90.93 (±0.22)	–	66.58 (±1.00)	–	62.15 (±1.35)	–
SI [2]	92.75 (±0.06)	1.81 (±0.21)	80.08 (±1.30)	5.54 (±1.30)	72.34 (±0.42)	8.26 (±0.64)
DER [3]	91.22 (±0.30)	4.63 (±0.26)	77.27 (±0.30)	9.31 (±0.09)	71.90 (±1.44)	8.36 (±2.06)
LUMP [6]	91.00 (±0.40)	2.92 (±0.53)	82.30 (±1.35)	4.71 (±1.52)	76.66 (±2.39)	3.54 (±1.04)
Cassle [7]	90.84 (±0.13)	2.29 (±0.23)	76.46 (±1.02)	3.05 (±0.87)	71.99 (±0.46)	3.34 (±0.52)
RM-SimSiam* (ours)	91.22 (±0.12)	4.15 (±0.18)	78.48 (±0.31)	4.09 (±0.99)	72.25 (±0.06)	4.51 (±0.04)
RM-SimSiam (ours)	**93.07 (±0.13)**	**1.36 (±0.10)**	**83.26 (±0.30)**	**2.73 (±0.42)**	**77.10 (±0.16)**	**2.67 (±0.01)**
MULTITASK	95.76 (±0.08)	–	86.31 (±0.38)	–	82.89 (±0.49)	–

the same as DER and LUMP, our RM-SimSiam beats all the other UCL methods and achieves new state-of-the-art results on all three benchmark datasets for UCL. This indicates that our proposed RM-SimSiam is indeed complementary to the rehearsal-based strategy and provides a new perspective to mitigate forgetting in UCL. **(3)** Our RM-SimSiam outperforms the latest rehearsal-based method LUMP [6] by 0.44% – 2.07% on accuracy and by 0.87% – 1.98% on forgetting, which provides direct evidence that our proposed rememory mechanism is crucial for learning the new task well while mitigating the forgetting under the UCL setting.

Table 2 shows the comparative results on the out-of-distribution (OOD) datasets. All UCL methods (with the same backbone ResNet18 [14]) are first trained on S-CIFAR-100, and then directly tested on the OOD datasets. Following [6], the OOD evaluation is performed on MNIST [20], Fashion-MNIST (FMNIST) [21], SVHN [22], CIFAR-10 [15], respectively. From Table 2, we can see that our RM-SimSiam clearly outperforms the state-of-the-art methods according to the average performance over all tasks. The obtained improvements on the OOD datasets show the superior generalization ability of our RM-SimSiam when unseen data distributions are encountered.

3.3. Ablation Study

To demonstrate the contribution of each key component (see Figure 1) of our full RM-SimSiam, we conduct ablation study on S-CIFAR-10. We take SimSiam [9] as the first baseline (denoted as Base). On the basis of Base, we add the Mixup strategy to form the second baseline (denoted as Base+Mixup). Further, we add other two key components including the rememory mechanism (RM) and the extra loss L_{hist} (Hist), which together make up our full RM-SimSiam (denoted as Base+Mixup+RM+Hist).

The ablation study results in Table 3 demonstrate that: **(1)** The Mixup strategy leads to improvements over Base (SimSiam), due to the use of the old data from the memory buffer. **(2)** Our rememory mechanism brings further improvements on both accuracy and forgetting (see Base+Mixup+RM vs. Base+Mixup). This suggests that our rememory mechanism is complementary to the rehearsal-based method based on Mixup. **(3)** When the extra loss L_{hist} is added, we can see significant improvements over Base+Mixup, which indicates that L_{hist} has important effect on the model performance. **(4)** The combination of RM and L_{hist} yields further improvements, showing their complementarity under the UCL setting. **(5)** Our full RM-SimSiam achieves significant improvements over Base+Mixup, which means that we have made sufficient contributions by devising new rememory mechanism and enhanced SimSiam-based contrastive loss for UCL.

Table 2. Comparison to the state-of-the-arts on the out-of-distribution (OOD) datasets.

IN-CLASS	S-CIFAR-100			
OUT-OF-CLASS	MNIST	FMNIST	SVHN	CIFAR-10
FINETUNE	85.99 (±0.86)	76.90 (±0.11)	50.09 (±1.41)	57.15 (± 0.96)
SI [2]	91.50 (±1.26)	80.57 (±0.93)	54.07 (±2.73)	60.55 (±2.54)
DER [3]	87.96 (±2.04)	76.21 (±0.63)	47.70 (±0.94)	56.26 (±0.16)
LUMP [6]	91.76 (±1.17)	81.61 (±0.45)	50.13 (±0.71)	63.00 (±0.53)
Cassle [7]	88.87 (±0.45)	81.30 (±0.45)	51.04 (±0.01)	59.46 (±1.62)
RM-SimSiam (ours)	**94.96 (±0.21)**	**83.29 (±0.19)**	**60.37 (±1.72)**	**69.16 (±0.17)**
MULTITASK	90.35 (±0.24)	81.11 (±1.86)	52.20 (±0.61)	70.19 (±0.15)

Table 3. Ablative results for our full RM-SimSiam on S-CIFAR-10. RM – the rememory mechanism; Hist – the extra contrastive loss L_{hist} defined based on the hist-module.

Method	S-CIFAR-10	
	acc (↑)	fg (↓)
Base (SimSiam)	90.16 (±0.24)	5.85 (±0.32)
Base+Mixup	90.40 (±0.18)	2.47 (±0.08)
Base+Mixup+RM	91.10 (±0.21)	1.67 (±0.41)
Base+Mixup+Hist	92.49 (±0.19)	1.96 (±0.26)
Base+Mixup+RM+Hist (full)	**93.07 (±0.13)**	**1.36 (±0.10)**

4. CONCLUSION

We propose a novel rememory-based method RM-SimSiam for unsupervised continual learning of image representation by storing and remembering the old knowledge with a data-free historical module to reduce the dependency on replay buffer. The proposed rememory mechanism and enhanced SimSiam-based contrastive loss largely improve the rememory ability of our RM-SimSiam. Extensive experiments demonstrate the effectiveness of our RM-SimSiam under the UCL setting. In future work, we plan to extend our method to other scenarios for unsupervised image representation, such as class incremental learning under the UCL setting.

5. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (62376274).

6. REFERENCES

[1] Mark Bishop Ring, *Continual Learning in Reinforcement Environments*, Ph.D. thesis, University of Texas at Austin, USA, 1994, UMI Order No. GAX95-06083.

[2] Friedemann Zenke, Ben Poole, and Surya Ganguli, "Continual learning through synaptic intelligence," in *International Conference on Machine Learning (ICML)*, 2017, pp. 3987–3995.

[3] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara, "Dark experience for general continual learning: a strong, simple baseline," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 15920–15930, 2020.

[4] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz, "Learning fast, learning slow: A general continual learning method based on complementary learning system," *arXiv preprint arXiv:2201.12604*, 2022.

[5] Zhiwei Lin, Yongtao Wang, and Hongxiang Lin, "Continual contrastive self-supervised learning for image classification," *arXiv preprint arXiv:2107.01776*, 2021.

[6] Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang, "Representational continuity for unsupervised continual learning," in *International Conference on Learning Representations (ICLR)*, 2021.

[7] Enrico Fini, Victor G Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal, "Self-supervised models are continual learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9621–9630.

[8] Michael McCloskey and Neal J Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of Learning and Motivation*, vol. 24, pp. 109–165. Elsevier, 1989.

[9] Xinlei Chen and Kaiming He, "Exploring simple siamese representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15750–15758.

[10] Russell A Poldrack, Jill Clark, EJet al Paré-Blagoev, Daphna Shohamy, J Creso Moyano, Catherine Myers, and Mark A Gluck, "Interactive memory systems in the human brain," *Nature*, vol. 414, no. 6863, pp. 546–550, 2001.

[11] Daphna Shohamy and Anthony D Wagner, "Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events," *Neuron*, vol. 60, no. 2, pp. 378–389, 2008.

[12] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3733–3742.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," *Handbook of Systemic Autoimmune Diseases*, vol. 1, no. 4, 2009.

[16] Arijit Banerjee and Vignesh Iyer, "Cs231n project report-tiny imagenet challenge," 2015.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[18] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[19] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[20] Yann LeCun, Corinna Cortes, and CJ Burges, "MNIST handwritten digit database. AT&T labs [online]," *yann. lecun. com/exdb/mnist*, 2010.

[21] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[22] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.