Instance and Category Supervision are Alternate Learners for Continual Learning

Xudong Tian[1,2,3], Zhizhong Zhang[1,3(✉)], Xin Tan[1,2,3], Jun Liu[4], Chengjie Wang[4], Yanyun Qu[5], Guannan Jiang[6], Yuan Xie[1,3(✉)]
[1]East China Normal University, [2]Shanghai Key Laboratory of Computer Software Testing & Evaluating, [3]Chongqing Institute of East China Normal University, [4]Tencent YouTu Lab, [5]Xiamen University, [6]Contemporary Amperex Technology Co., Limited
txd51194501066@gmail, {zzzhang,xtan}@cs.ecnu.edu.cn, junsenselee@gmail.com
jasoncjwang@tencent.com, yyqu@xmu.edu.cn, jianggn@catl.com, xieyuan8589@foxmail.com

Abstract

Continual Learning (CL) is the constant development of complex behaviors by building upon previously acquired skills. Yet, current CL algorithms tend to incur class-level forgetting as the label information is often quickly overwritten by new knowledge. This motivates attempts to mine instance-level discrimination by resorting to recent self-supervised learning (SSL) techniques. However, previous works have pointed out that the self-supervised learning objective is essentially a trade-off between invariance to distortion and preserving sample information, which seriously hinders the unleashing of instance-level discrimination.

In this work, we reformulate SSL from the information-theoretic perspective by disentangling the goal of instance-level discrimination, and tackle the trade-off to promote compact representations with maximally preserved invariance to distortion. On this basis, we develop a novel alternate learning paradigm to enjoy the complementary merits of instance-level and category-level supervision, which yields improved robustness against forgetting and better adaptation to each task. To verify the proposed method, we conduct extensive experiments on four different benchmarks using both class-incremental and task-incremental settings, where the leap in performance and thorough ablation studies demonstrate the efficacy and efficiency of our modeling strategy.

1. Introduction

Humans learn from visual inputs with everchanging scenarios, both rapidly and flexibly absorbing new knowledge with constantly emerging concepts, and robustly accumulating previously acquired experiences. Modeling such powerful capability is the central target of continual learning (CL), and would be of substantial utility in real-world computer vision settings [44].

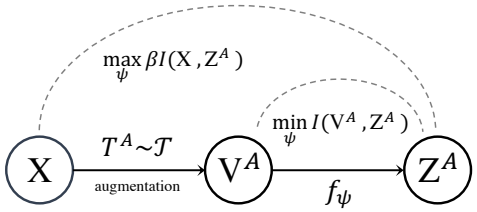


Figure 1: Illustration of the trade-off that is widely embodied in recent self-supervised learning techniques. X denotes the given sample, V^A and Z^A stand for the augmented view and the corresponding embedding obtained from distortion $T^A \sim \mathcal{T}$ and network f_ψ, respectively.

To this end, a variety of CL algorithms [7, 8, 19, 28, 40] have been developed to get rid of the requirement of *i.i.d* samples, and attempt to alleviate catastrophic forgetting [23] when learning a continuum of training data. Broadly speaking, studies on this topic can be categorized into three schools: (i) rehearsal-based methods [4, 14, 33], which store samples in raw or generative format, and replay them to alleviate forgetting; (ii) regularization-based methods [32, 36] that impose extra constraints to prediction, gradient, *etc.* to consolidate previously learned contents; (iii) parameter isolation strategies [13, 22] that dedicate or mask a part of the model parameters for the training of each task.

Albeit the differences, the typical solution is to utilize the label, *i.e.*, category-level supervision to acquire new knowledge while accumulating previously learned contents. Although this usually leads to fast adaptation, such strategy, *i.e.*, supervised learning (SL), is prone to incur class-level

forgetting and quickly fades out with the label information overwritten by new concepts [23, 34]. Moreover, they also tend to produce overfitted and biased models when only a small amount of datum (*e.g.*, incremental tasks, or samples stored in exemplar memory) is accessible for training.

The deficiencies mentioned above motivate a trend in CL community to resort to instance-level discrimination to alleviate forgetting. For example, [21] directly applies SSL constraints to a pre-trained model, [25] deploys a parallel self-supervised branch in addition to a supervised partner. Unfortunately, as indicated in BarlowTwins [41], the objective of SSL is essentially a trade-off between preserving sample information and being invariant to distortions, but both of which are necessary and beneficial for continual learning. As a consequence, continual learners equipped with such strategy can barely achieve instance-level discrimination, and are heavily dependent on an extra supervised partner. The above factors seriously limit the effectiveness of the self-supervised continual learner and have a detrimental effect on the performance of the entire CL framework, especially in task-agnostic CL settings [38].

In this work, we tackle the trade-off by reformulating the SSL objective from the information-theoretic perspective. More specifically, we first disentangle the principal target of instance-level discrimination into two terms, *i.e.*, (i) maximizing sample information without intensifying variation caused by distortion; (ii) promoting invariance to yield compact representations. On this basis, our SSL strategy exhibits superiority in preserving instance-level discrimination, and yields improved robustness against forgetting.

To enjoy both complementary merits of category-level and instance-level supervision, we develop a novel paradigm for continua learning. Concretely, it includes two updates to the continual learner, *i.e.*, an "*inner-loop*" which alternately conducts SSL and SL, and an "outer-loop" which uses a momentum update to accumulate knowledge from previous tasks. In such cases, SSL serves as a pre-training procedure and maintains stability, while SL is utilized to generate task-specific parameters for plasticity.

To validate the proposed method, we conduct extensive experiments on four benchmarks, including CIFAR-100 [16], Tiny-ImageNet [17], ImageNet-100 and ImageNet-1K [5], and provide a comprehensive ablation on each component to show the qualitative characteristics. Our contributions can be summarized as follows:

(i) We reformulate the SSL objective by disentangling it into two terms, which promote invariance against distortion while simultaneously producing compact representations.

(ii) We design a novel alternate paradigm for continual learning, which fully exploits the complementary advantages of both category-level and instance-level supervision, demonstrating significant superiority in achieving both stability and plasticity.

(iii) The leap in performance compared with all competitors on various benchmarks demonstrates its efficacy, while substantial qualitative evidence verifies each of our designs.

2. Preliminary and Related Work

As analyzed in [21, 44], instance-level discriminative learning is helpful to alleviate forgetting, which motivated a series of continual learning works [20, 25, 39] to embody self-supervised training techniques. In this section, we will give analyses on instance-level supervision, and show the major bottleneck of current self-supervised continual learners from an information-theoretic perspective.

2.1. Connections between SSL and CL

Given x and y as an input sample and its ground-truth label, respectively. As analyzed in [25], supervised learner aims to purify $I(x;y)$[1] for fast adaptation to new knowledge, where $I(x;y)$ denotes the label information contained in x, and is identical among the corresponding category. But as a by-product, it would also incur class-level forgetting when $I(x;y)$ is overwritten by new concepts $\hat{y}$ [34].

On the other hand, the consensus in SSL posits that good representations should be associated with instance-level discrimination, which encourages to produce representation z to (i) be maximally informative with regard to the sample x itself (*i.e.*, maximizing $I(x;z)$); (ii) simultaneously maintaining invariant to disturbance like image distortions (*i.e.*, minimizing $I(v^A;z)$ with v^A defined as the distorted viewpoint). Notably, since the learning objective of instance-level discrimination would not change with continual learning, SSL demonstrates better robustness against forgetting [1, 21].

2.2. Trade-off in Current SSL Strategy

Given sample $x_i \in X$, common practice [3, 41] in SSL first generates two augmented viewpoints x_i^A and x_i^B, then forwards both of them to a backbone network and obtains embedding z_i^A and z_i^B. Afterward, constraints are imposed to preserve instance-level discrimination for sample x_i. For example, the most widely adopted contrastive learning [3, 11] is formed to minimize $\frac{d(z_i^A, z_i^B)}{\sum_{j=1}^{N} d(z_i^A, z_j)}$ with $d(\cdot,\cdot)$ and z_j defined as a similarity metric and embedding obtained from $x_j (\forall j \neq i)$, respectively.

However, as analyzed in [41], the commonly adopted SSL strategy for continual learning is essentially a trade-off between the desiderata of preserving information and being invariant to distortions. More specifically, it can be instantiated as (please see Fig. 1 for graphical illustration):

$$\mathcal{L}_{SSL} \triangleq \underbrace{I\left(V^A; Z^A\right)}_{\text{variation}} - \underbrace{\beta I\left(Z^A; X\right)}_{\text{sample info}}. \tag{1}$$

[1] Mutual information between x and y.

In particular, larger β promotes invariance against the distortion but at the cost of losing more sample information, while smaller β encourages the representation to be informative with respect to the sample X itself but may result in being vulnerable to the distortion.

As a consequence, the continual learner with Eq. (1) meets the trade-off challenge and cannot fully realize the power of SSL when dealing with sequentially arrived tasks [38]. Next, we introduce an improved SSL strategy that effectively maintains stability by tackling the trade-off in Eq. (1). Moreover, a new alternate learning paradigm is developed to enjoy the complementary merits of both instance-level and category-level supervision.

3. Method

3.1. Problem Setup

Continual learning is defined as training machine learning models on a sequence of tasks. Formally, consider the sequence of tasks as $\mathcal{D} = \{\mathcal{D}^0, \mathcal{D}^1, ..., \mathcal{D}^T\}$, where the t-th task $\mathcal{D}^t = \{(x_i^t, y_i^t)_{i=1}^{n_t}\}$ includes the input sample $x_i^t \in \mathcal{X}$ in total of n_t and its ground-truth label $y_i^t \in \mathcal{Y}$. Note $\forall i, j$, $0 \leq i, j \leq T$ and $i \neq j$, $\mathcal{D}^i \cap \mathcal{D}^j = \emptyset$, *i.e.*, there are no overlaps between different tasks. In contrast to classical training taxonomy where all training samples are accessible, a certain budget (*e.g.*, 1,000 exemplars) is determined for the memory $\mathcal{M}$ such that a subset of observed data from previous tasks can be stored by $\mathcal{M} = \{(x_i^t, y_i^t)_{i=1}^{N}\}$. For the first task, the training data merely consists of $\mathcal{D}^0$. Afterward, the training set is formed as a merged sub-dataset and includes data from the current task and the memory, *i.e.*, $\mathcal{D}^t \cup \mathcal{M}$.

Our model $\Omega = \{\theta, \phi, \psi\}$ is composed of an encoder parameterized by θ to produce observation $v = f(x; \theta)$ from the input sample x, a classifier parameterized by ϕ to make prediction $\hat{y} = f(v; \phi)$ based on the observation, and a projector parameterized by ψ to generate projection $z = f(v; \psi)$ used for self-supervised learning. At time t, our goal is to update Ω^t from Ω^{t-1} based on $\mathcal{D}^t \cup \mathcal{M}$.

3.2. An Improved Self-Supervised Learning Strategy for Continual Learning

To improve the robustness against forgetting, we next present an improved SSL strategy to tackle the trade-off in Eq. (1), which effectively preserves sample information while being invariant to distortion.

Given an augmented view $x^A = T^A(x)$ of x, our model outputs the observation $v^A = f(x^A; \theta)$ and embedding $z^A = f(v^A; \psi)$ from the encoder and projector, respectively. On this basis, we first show the following factorization based on the chain rule [9]:

$$I(v^A; z^A) = I(z^A; z^B) + I(v^A; z^A | z^B), \tag{2}$$

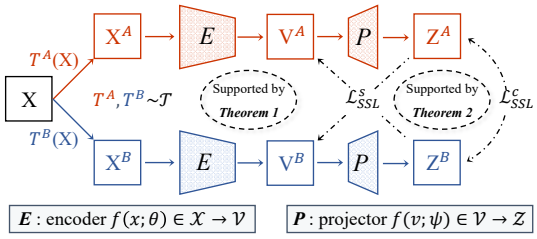


Figure 2: Graphical illustration of the self-supervised learning procedure, where $\mathcal{T}$ is a distribution of data augmentation, and X denotes the given sample.

where $I(z^A; z^B)$ denotes the invariance when applying data augmentation, *i.e.*, sample information, and $I(v^A; z^A | z^B)$ represents information irrelevant to the sample itself, *i.e.*, variation caused by the distortion T^A.

Recall Eq. (1), which intends to minimize $I(v^A; z^A)$. Eq. (2) suggests that reduction to $I(v^A; z^A)$ jointly decreases $I(z^A; z^B)$, which is necessary to unleash instance-level discrimination. As a result, the min-max game makes both targets entangled and incurs the trade-off. On the contrary, Eq. (2) formulates a disentangled factorization by dividing $I(v^A; z^A)$ into two terms. Minimizing $I(v^A; z^A | z^B)$ eliminates variation with the invariance unharmed, and maximizing $I(z^A; z^B)$ promotes the preservation of the sample information without intensifying the variation.

To maximize the invariant information in z^A, we first utilize the information processing inequality to show:

$$I(z^A; z^B) < I(z^A; v^B). \tag{3}$$

Here, z^B is obtained by our projector head, and therefore the inequality holds. Note Eq. (3) suggests an improved solution for z^A by reformulating the preservation of sample information as maximizing $I(z^A; v^B)$, which is consistent with $I(z^A; z^B)$ but promises more potential with the help of our Theorem 2. However, optimization to mutual information is notoriously difficult, especially when dealing with high-dimensional variables like z^A. To address this, we introduce the following theory for promoting invariance.

Theorem 1. *Given v^A, v^B as the observation of x from different augmentations, $I(z^A; v^B)$ is maximally preserved when (vice versa for z^B):*

$$D_{KL}\left[p(v^B|v^A)||p(v^B|z^A)\right] = 0$$

where D_{KL} denotes the KL-divergence, and $p(\cdot|\cdot)$ represents conditional distribution.

Please refer to the supplementary material for detailed proof and implementation. Specifically, Theorem 1 promises the self-supervised learner to be capable of main-

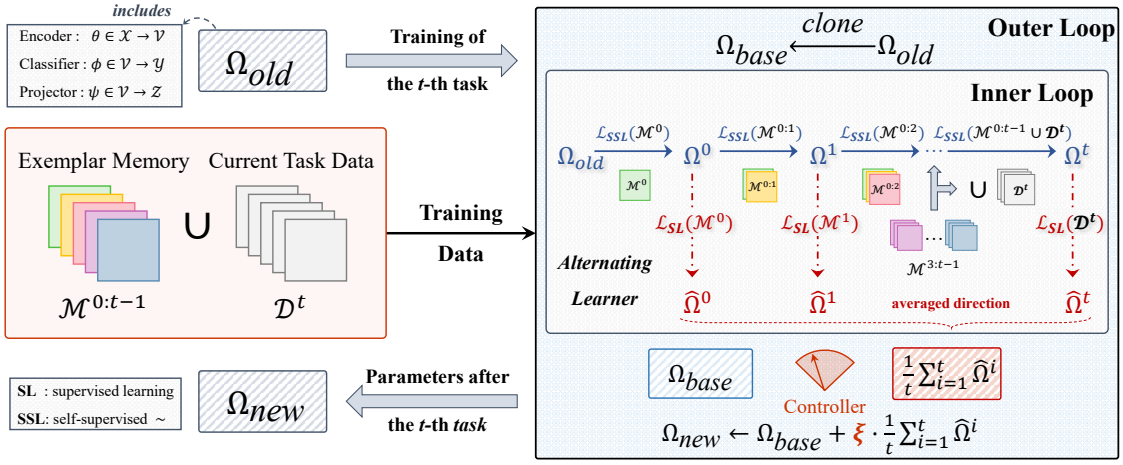


Figure 3: Training paradigm of our alternate learner, where self-supervised and supervised learning procedures are alternately conducted, and the new parameters are optimized in a momentum manner.

taining stability by optimizing the following objective:

$$\mathcal{L}^s_{SSL} = \mathbb{E}_{v\sim f_\theta(x)}\mathbb{E}_{z\sim f_\psi(v)} D_{KL}\left[p(v^B|v^A)||p(v^B|z^A)\right]. \tag{4}$$

Next, to minimize the variation $I(v^A; z^A|z^B)$, we present the following theory.

Theorem 2. *Given v^A, v^B as the observation of x from different augmentations, and z^A, z^B are the corresponding representations, we have:*

$$D_{JS}\left[p(z^A|v^A)||p(z^B|v^B)\right] = 0 \Rightarrow \begin{cases} I(v^A; z^A|z^B) = 0, \\ I(v^B; z^B|z^A) = 0. \end{cases}$$

where D_{JS} is the Jensen–Shannon divergence.

Please refer to the supplementary material for proof and implementation details. Formally, Theorem 2 formulates the following objective to eliminate the variation:

$$\mathcal{L}^c_{SSL} = \mathbb{E}_{v\sim f_\theta(x)}\mathbb{E}_{z\sim f_\psi(v)} D_{JS}\left[p(z^A|v^A)||p(z^B|v^B)\right]. \tag{5}$$

Notably, Eq. (5) is disentangled from the sample information, which explicitly removes detriments to the invariance. On this basis, z^A is facilitated to mine more invariant clues from v^B without intensifying the variation, promising more potential to optimize $I(z^A; v^B)$ instead of $I(z^A; z^B)$.

Combining with Eq. (4), the overall loss function of our SSL strategy can be given as:

$$\mathcal{L}_{SSL} = \alpha_s \cdot \mathcal{L}^s_{SSL} + \alpha_c \cdot \mathcal{L}^c_{SSL}, \tag{6}$$

which promotes instance discrimination while simultaneously encouraging compact embedding. Note α_s and α_c define the weight and are fixed to 0.65 and 0.25 by default.

Discussion. The trade-off in common SSL approaches inevitably compromises stability [44, 21, 25], resulting in inferior performance [38]. By comparison, our strategy provides an analytical solution to keeping instance discrimination yet yields compact representations. On this basis, it tackles the trade-off and demonstrates significant superiority in alleviating forgetting.

Next, we present a novel alternate learning paradigm to enjoy the complementary merits of instance-level and category-level supervision.

3.3. Alternate Learner

As shown in Fig. 3, our strategy includes two updates to the continual learner, *i.e.*, an "*inner-loop*" which alternately conducts SSL and SL to generate task-specific parameters, and an "outer-loop" which uses a momentum update to accumulate knowledge from previous tasks.

Our motivations to design an alternate paradigm are twofold. First, as analyzed in Sec. 2.1, self-supervised learners encourage instance-level discrimination and are more robust against forgetting, while supervised learners focus on label information which can better and rapidly adapt to each task. Thus the alternate training strategy can enjoy both merits of SSL and SL. Second, the widely adopted empirical evidence suggests SSL pre-training usually leads to a better convergence for supervised learners, thus suggesting an alternate training style would be favorable.

Inner Loop. For the first task ($t = 0$), the model sequentially conducts SSL and SL on $\mathcal{D}^0$ to obtain Ω^0 and $\widehat{\Omega}^0$, respectively. For the t-th task ($t \geq 1$), where the training set is the union of $\mathcal{M}^{0:t-1}$ (exemplars of previous tasks) and $\mathcal{D}^t$ (accessible data from current task), samples are first partitioned according to their task ID and generates a series of subsets for training task-specific parameters (see $\mathcal{M}^0$, $\mathcal{M}^{0:1}$, *etc.* in Fig. 3). Note the task identifier stored in

the memory does not violate the class-incremental protocol, and for the t-th task by giving the parameters Ω_{old} from previous tasks the training procedure can be specified as:

• The learner Ω_{old} starts from performing SSL on $\mathcal{M}^0$, and obtains Ω^0.

• Afterward, Ω^0 is trained to minimize binary cross entropy (*i.e.*, $\mathcal{L}_{SL}$) with ground-truth label on $\mathcal{M}^0$ to generate the task-specific parameters $\widehat{\Omega}^0$.

• Next, the learner starts from Ω^0 and carries on SSL with $\mathcal{M}^{0:1}$ to obtain Ω^1.

• Similarly, we have $\widehat{\Omega}^1$ and Ω^2 after performing SL and SSL on Ω^1 with $\mathcal{M}^1$ and $\mathcal{M}^{0:2}$, respectively.

• Such alternating training proceeds to the t-th task and generates Ω^t, $\widehat{\Omega}^t$ using $\mathcal{M}^{0:t-1} \cup \mathcal{D}^t$ and $\mathcal{D}^t$ respectively.

Please refer to Fig. 3 and Algorithm 1 (2^{nd}-19^{th} line) for graphical and mathematical illustration of the inner loop. Although each task requires the model to be trained from Ω^0, the training data corresponding to previous tasks only comes from the memory, which stores a tiny part of samples, thus leads to acceptable complexity. Detailed analysis are provided in Sec. 4.3.

Discussion. Compared with [21] which directly applies Eq. (1) to IL or [25] that rigidly combines SSL and SL, our strategy fully exploits the advantages of both self-supervised and supervised learners. Specifically, $\{\Omega^i\}_{i=0}^t$ accumulates knowledge throughout the incremental procedure to preserve instance discrimination and make the model not biased towards any of $\mathcal{M}^0, ..., \mathcal{M}^{t-1}, \mathcal{D}^t$. Meanwhile, by alternately training with the ground-truth label, $\{\widehat{\Omega}^i\}_{i=0}^t$ generates a series of task-specific parameters with better adaptation to each task.

Outer Loop. To incorporate knowledge from different tasks, we next integrate the task-specific parameters generated during the inner loop to formulate a generic model. Concretely, the gradient update of the outer loop is the combination of $\left(\Omega_{base} - \widehat{\Omega}^i\right)$ for all $\{\widehat{\Omega}^i\}_{i=0}^t$, where Ω_{base} denote the clone of initial parameters Ω_{old}. On this basis, the meta-model is optimized towards the average direction of all task-specific updates, and thus yields better generalization [10, 24]. Formally, the outer loop update is formed as:

$$\Omega_{new} \leftarrow \Omega_{base} + \xi \cdot \frac{1}{t+1} \sum_{i=0}^{t} \widehat{\Omega}^i. \tag{7}$$

ξ is a momentum-based controller [11, 24] defined as $\xi = \exp(\frac{\eta}{t+1})$, where η denotes a constant decay rate. Practically, the controller encourages fast adaption in the beginning and gradually focuses on maintaining stability with the increase of tasks. More illustrations for the outer loop could be found in Fig. 3 and Algorithm 1.

Discussion. In essence, the outer loop serves as a gradient-based regularization, and ensures the model to be optimized towards the direction that is beneficial to all tasks. Particularly, this coincides with the task-shared gradient constraints in [32, 36]. Moreover, such a paradigm only moves the feature space towards the optimal manifold of each task, while encouraging the classifier to be as close as possible to the corresponding task-specific optimal solution.

Algorithm 1 Alternate Learner

Input: initial parameters Ω_{old}, training data $\mathcal{M}^{0:t-1} \cup \mathcal{D}^t$, a distribution of data augmentations $\mathcal{T}$

$\Omega_{base} \leftarrow \Omega_{old}$
for $i \leftarrow 0$ to t **do** ▷ ***inner loop starts***
  $\Omega^i \leftarrow \Omega_{old}$ **if** $i = 0$ **else** $\Omega^i \leftarrow \Omega^{i-1}$
  $\widehat{\mathcal{D}}^i_{SSL} \leftarrow \mathcal{M}^{0:i}$ **if** $i < t$ **else** $\widehat{\mathcal{D}}^i_{SSL} \leftarrow \mathcal{M}^{0:i} \cup \mathcal{D}^t$
  for $e \in e_{SSL}$ **do** ▷ ***SSL procedure***
    sample $X = \{x_j\}_{j=1}^N \sim \widehat{\mathcal{D}}^i_{SSL}$
    obtain $T^A, T^B \sim \mathcal{T}$
    $X^A \leftarrow T^A(X)$, $X^B \leftarrow T^B(X)$
    $loss \leftarrow \frac{1}{N} \sum_{j=0}^{N} \mathcal{L}_{SSL}(f_\theta^A(X^A), f_\theta^B(X^B))$
    $\Omega^i \leftarrow$ Optimizer$(\Omega^i, loss)$
  end for
  $\widehat{\mathcal{D}}^i_{SL} \leftarrow \mathcal{M}^i$ **if** $i < t$ **else** $\widehat{\mathcal{D}}^i_{SL} \leftarrow \mathcal{D}^t$
  $\widehat{\Omega}^i \leftarrow \Omega^i$
  while not convergent **do** ▷ ***SL procedure***
    sample $\mathcal{B} = \{(x_j^i, y_j^i)\}_{j=1}^M \sim \widehat{\mathcal{D}}^i_{SL}$
    $loss \leftarrow \frac{1}{M} \sum_{j=0}^{M} \mathcal{L}_{SL}(y_j^i, \widehat{\Omega}^i(x_j^i))$
    $\widehat{\Omega}^i \leftarrow$ Optimizer$(\widehat{\Omega}^i, loss)$
  end while
end for
$\Omega_{new} \leftarrow \Omega_{base} + \xi \cdot \frac{1}{t} \sum_{i=1}^{t} \widehat{\Omega}^i$ ▷ ***outer loop update***

Output: Ω_{new}

4. Experiments

In this section, we investigate our approach via comprehensive experiments and sensitivities on a variety of public datasets and demonstrate its ability to facilitate both stability and plasticity.

4.1. Experimental settings

Datasets. The CIFAR-100 [16] dataset is composed of 100 categories and each of them includes 500 training and 100 testing samples with the same size of 32×32. ImageNet-1K [5] contains over 1.2 million images with 1000 different classes. ImageNet-100 [5, 28] is built by selecting a subset of ImageNet-1K and contains 100 categories, each of which is associated with over 1000 samples of 224×224 size. TinyImageNet [17] is another variant of ImageNet-1000 that consists of 200 classes with the image size of 64×64. In the following experiments, we adopt both class-incremental and task-incremental protocols, and utilize 10/20-split (randomly and evenly splitting the bench-

Table 1: Comparison on CIFAR-100, Image-100, and ImageNet-1K using class-incremental protocol, averaged across 3 trials. Average accuracy (avg. acc.) and last accuracy (last acc.) are adopted for evaluation, red and blue values denote the best and secondary performance.

Method	Venue	CIFAR-100		ImageNet-100 10-split		ImageNet-1000 10-split	
		acc. (10-split)	acc. (20-split)	avg. acc.	last acc.	avg. acc.	last acc.
iCaRL [28]	CVPR'17	65.27±1.02	61.20±0.83	-	-	-	-
UCIR [12]	CVPR'19	58.66±0.71	58.17±0.30	-	-	-	-
BiC [12]	CVPR'19	68.80±1.20	66.48±0.32	-	-	-	-
PODNet [7]	ECCV'20	63.19±1.16		74.33	-	64.12	-
AANet [19]	CVPR'21	64.31±0.90	-	75.58±0.74	-	64.85±0.53	-
PASS [42]	CVPR'21	61.84	58.09	61.80	-	-	-
ELI [15]	CVPR'22	61.72	57.65	55.47	-	-	-
FOSTER [35]	ECCV'22	72.90	70.65	77.75	-	68.34	-
DiversMem [33]	CVPR'22	66.47	-	76.76	-	-	-
FCIL [6]	CVPR'22	66.9	-	57.0	-	-	-
SSRE [43]	CVPR'22	65.04	61.70	67.69	-	-	-
DER (w/o P) [40]	CVPR'21	75.36±0.36	74.09±0.33	77.18	**66.70**	68.84	60.16
DyTox+ [8]	CVPR'22	**76.74±1.08**	**76.25±0.30**	**77.62**	65.94	**73.21**	**64.56**
Ours	-	**79.16±0.31**	**77.58±0.29**	**80.11**	**69.63**	**73.25**	**65.07**

marks with 10/20 disjoint task) for comparisons.

Evaluation Metrics. For all datasets, we follow [8, 21, 40] and adopt average accuracy and average forgetting for comparison. Average accuracy (**avg. acc.**) is the average test accuracy of all the tasks completed until the continual learning of task t. Average forgetting (**avg. fgt.**) is the average performance decrease of each task between its maximum accuracy and accuracy at the completion of training: $\frac{1}{T-1}\sum_{i=1}^{T-1}\max_{t\in\{1,\dots,T\}}(a_{t,i}-a_{T,i})$, where $a_{i,t}$ denotes the test accuracy of task i after learning the t-th task.

Implementation Details. For all benchmarks, we follow [40] and adopt ResNet-18 as our encoder, and implement the projector with multi-layer perceptrons of 2 hidden ReLU units of size 512 and 256 respectively with an output of size 128 that produces embeddings z^A and z^B. For training, all experiments are optimized by RAdam [18] with an initial learning rate of 10^{-2} for 100 epochs, and the self-supervised learning iteration e_{SSL} is fixed to 15 for all tasks. The batch size of CIFAR-100 and ImageNet-100 is set to 128, and it increases to 256 and 512 for TinyImageNet and ImageNet-1K, respectively. The learning rate decays 5 times after training 30, 60, and 90 epochs. Following [27, 28, 40], we keep a fixed memory size of $2,000$ exemplars for all settings of CIFAR-100 and ImageNet-100, 4K for TinyImageNet, and 20K for ImageNet-1K. All the models are trained on NVIDIA RTX A6000 GPUs.

4.2. Comparison with State-of-the-art Methods

Class-incremental Learning. Fig. 4 and Tab. 1 summarize the results obtained by using class-incremental protocol. We can see that our method consistently outperforms all competitors by a sizable margin at different incremental splits on all benchmarks. When continually learning 10 tasks, our method surpasses the transformer-based DyTox+ [8] by 2.42%, 2.49% on CIFAR-100 and ImageNet-100, respectively. On the other hand, the margin evidently increases when compared with CNN-based state-of-the-arts, *e.g.*, outperforming FOSTER [35] and DER[40] by 4.41%, 4.91% on the large-scale ImageNet-1K. Similar conclusions can be given with 20-split results, where our approach achieves 3.49% gain compared with [40].

Table 2: Comparison on CIFAR-100 and TinyImageNet (T-IMN) utilizing the task-incremental setting, averaged across 3 trials. Red and blue indicate methods with the best and secondary performance (average accuracy).

Methods	Venue	CIFAR-100		T-IMN
		10-split	20-split	20-split
AGEM [2]	ICLR'19	40.38±0.30	42.39±0.42	28.38±0.15
ER [29]	ICLR'19	63.39±1.37	-	45.50±0.61
ASER [30]	AAAI'21	58.91±1.76	62.47±1.82	46.02±0.82
CTN [26]	ICLR'21	-	67.65±0.43	-
CVT [37]	ECCV'22	65.86±1.24	75.76±0.93	48.50±0.88
UCL [21]	ICLR'22	68.42±1.17†	**82.30±1.35**	**76.66±2.39**
MetaSP [31]	NeurIPS'22	**78.27±0.89**	-	-
Ours	-	**80.04±1.69**	**86.45±1.41**	**82.81±1.95**

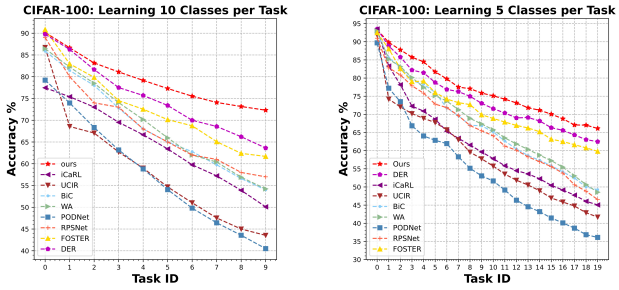
† Results obtained by using the official implementation.

Task-incremental Learning. As shown in Tab. 2, the results obtained in this task-incremental setting with our approach constitute a leap in performance compared to all other competitors. On the other hand, the results (comparing our method and UCL [21] with all others) also suggest that self-supervised training appears to be favorable for learning sequential tasks compared with vanilla supervised opponents, which compiles with our analysis in Sec. 2.1.

4.3. Ablation Study

In this section, we validate each specific design in our approach via a thorough ablation study.

We first clarify various settings in Tab. 3. "SSL" and "SL" denote the continual learner trained with self-supervised and supervised strategies, respectively. "DN", *i.e.*, DualNet strategy, denotes both SSL and SL are deployed, and are incorporated as [25], where the self-supervised learner only processes data from memory and



(a) Accuracy (10-split) (b) Accuracy (20-split)

Figure 4: Classification accuracy on CIFAR-100, with 10 and 20 tasks for the left and right, respectively.

Table 3: Ablation study on CIFAR-100 (C100) and Tiny-ImageNet (T-IMN) under task-incremental setting. "Acc" and "Fgt" denote average accuracy and average forgetting, respectively.

Setting / Method	C100 10s		C100 20s		T-IMN 20s	
	Acc	Fgt	Acc	Fgt	Acc	Fgt
SSL	68.4	5.08	82.30	4.71	76.66	3.54
SSL+AL$_{out}$	68.9	4.71	82.41	4.49	76.91	3.40
SSL+SL	69.1	5.94	83.16	6.30	77.52	5.29
SSL+SL+DN	70.2	4.78	83.99	4.53	78.41	3.89
SSL+SL+AL$_{in}$	75.8	4.57	84.67	4.48	80.93	3.84
SSL+SL+AL	**77.5**	**4.09**	**85.21**	**4.39**	**81.43**	**3.68**
SSL+SL+$\widehat{\text{AL}}$	**80.0**	**3.56**	**86.45**	**4.21**	**82.81**	**3.31**

the supervised learner handles data from both current and previous tasks. "AL" and "$\widehat{\text{AL}}$" represent our alternate learner trained with and without Eq. (4) and Eq. (5), respectively. "AL$_{out}$" and "AL$_{in}$" denote only the outer or inner loop is applied.

Ablation on SSL and SL. As shown in Tab. 3 (see the first and third row), an extra supervised branch can bring minor improvement to the original self-supervised learner (*e.g.*, 0.86% increment of avg. acc. on TinyImageNet), but intensifies forgetting in the same time.

Ablation on the alternate learner. The 3$^{\text{rd}}$ and 4$^{\text{th}}$ row in Tab. 3 indicate incorporating the self-supervised learner with a supervised partner would be beneficial. Meanwhile, it also suggests that a straightforward or rigid combination results in a negligible promotion (*e.g.*, 0.89% gain on Tiny-ImageNet). By contrast, our alternate learner fully exploits the advantage and complementarity of both SSL and SL, hence significantly boosting the performance (compare the 3$^{\text{rd}}$ and 6$^{\text{th}}$ row) by 8.4% on CIFAR-100 (10-split).

To further provide qualitative evidences, we plot the 2D projection of the embedding space (obtained by using CIFAR-100 with 20-split setting) on Fig. 5. As is illustrated, we have the following observations:

(1) From Ω^1 to Ω^2, the embedding space remains relatively stable, and all clusters are compact after SSL. Though there exists shifted classes, we notice that such phenomenon

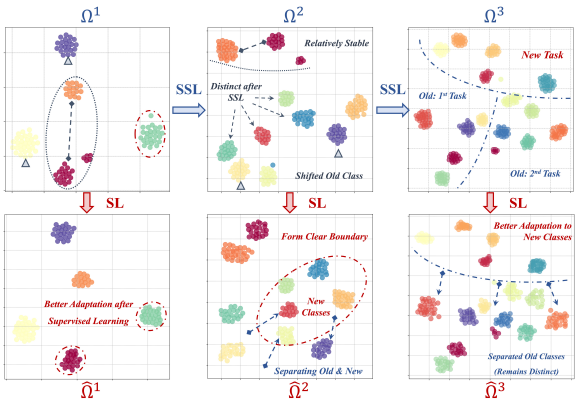


Figure 5: 2D Projection of the embedding space by using t-SNE. The results are obtained from our alternate learner by learning 5 classes per task on CIFAR-100. Different colors are used to represent different classes.

is quite common when encountering the first incremental task (see $0 \rightarrow 1$ in Fig. 4).

(2) From Ω^2 to $\widehat{\Omega}^2$, SL demonstrates better adaptation to the new classes, and forms a clear boundary between embeddings from current and previous tasks. More importantly, there are no apparent drifts to the old classes except being separated from new ones. Such phenomenon suggests incorporating SL sequentially with SSL can better fit new task without compromising previously learned classes.

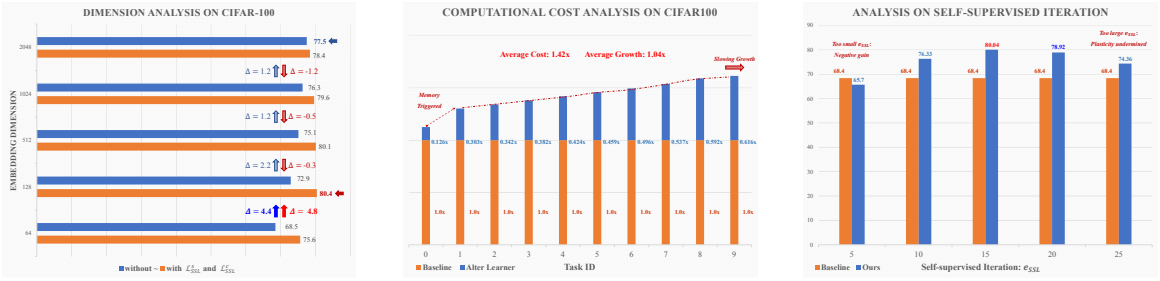
(3) From Ω^2 to Ω^3, classes from the same task gather in a specific space, while embeddings from different tasks are distributed to different areas. Meanwhile, all clusters are distinct and concentrate, revealing the capability to preserve instance discrimination and maintain stability.

(4) From Ω^3 to $\widehat{\Omega}^3$, similar as (2), the embeddings from current and previous tasks are clearly separated without causing drastic shifts to the old classes.

Based on the above analysis, we conclude that, by alternately conducting SSL and SL, our alternate learner can effectively maintain both stability and plasticity.

Ablation on the inner and outer loop. Since "AL$_{out}$" applies only the outer loop and removes the entire inner loop, it is essentially equivalent to updating the SSL branch by incorporating parameters obtained from previous steps (see blue markers $\{\Omega^i\}_{i=1}^t$ in Fig. 3) in a momentum fashion. On the other hand, "AL$_{in}$" represents only the inner loop is adopted, *i.e.*, Ω_{new} is obtained from $\widehat{\Omega}^t$.

By comparing the first two rows in Tab. 3, we observe the direct application of the outer loop can alleviate forgetting to some extent, and slightly improve the accuracy. However, the improvement is negligible due to the limited ability of a self-supervised learner to fit each task. On the other hand, "AL$_{in}$" demonstrates remarkable efficacy by providing a dramatic gain of 7.4%@Acc on CIFAR-100 10-split, while evidently reducing forgetting at the same time (see the 5^{th} and 1^{st} row in Tab. 3). Moreover, compared



(a) Dimension Analysis (b) Computation Analysis (c) SSL iteration Analysis

Figure 6: (a): Evaluation of different dimensions of the embedding. (b): Analysis on the extra computation brought by our alternate learner. (c): Performance of our method when adopting different self-supervised iteration e_{SSL}. All experiments are conducted on CIFAR-100 using 10-split protocol.

with the rigid combinations with an SL partner, "AL$_{in}$" fully exploits the advantages of both SSL and SL, which constitutes a leap in performance, *e.g.*, 6.7%, 5.6%@Acc on CIFAR-100 10-split over the 3^{rd} and 4^{th} row in Tab. 3.

Ablation on our SSL strategies. In addition to the above qualitative evidence, we can spot a considerable gain provided by our SSL strategy to the performance (see the last two rows in Tab. 3). More importantly, as illustrated in Fig. 6a, we give the following analysis:

(1) Contrast to [21, 41], our strategy does not require high-dimensional outputs to comply with the trade-off, and evidently surpasses the opponents with much fewer output channels (*i.e.*, 128 *vs.* 2048), demonstrating superiority in both efficacy and efficiency.

(2) By removing Eq. (4) and Eq. (5), the accuracy (see blue markers in Fig. 6a) slowly grows with the rapidly increased dimension. Such phenomenon is consistent with previous analysis of the trade-off Eq. (1), *i.e.*, conserving invariant information at the cost of introducing overwhelming variation resulted from distortion.

(3) By comparison, our strategy performs optimally with a rather compact embedding. We conjecture that extreme reduction to the output dimension might result in insufficiency to preserve discriminative cues. On the other hand, redundant channels appear to be more prone to introduce task-irrelevant nuisances, and thus cause degradation.

Complexity Analysis. The computational cost brought by our alternate learner is visualized in Fig. 6b. Suppose the complexity of performing supervised and self-supervised training on the exemplar memory is $\mathcal{O}_{\text{SL}}(|\mathcal{M}|)$ and $\mathcal{O}_{\text{SSL}}(|\mathcal{M}|)$, respectively (vice versa for current data $\mathcal{D}^t$). The first task ($t = 0$) leads to $\mathcal{O}_{\text{SL}}(|\mathcal{D}^t|)$ more computation. The successive tasks ($t \geq 1$) brought $\mathcal{O}_{\text{SL}}(|\mathcal{M}|) + \mathcal{O}_{\text{SL}}(|\mathcal{M}| \cdot \sum_{i=1}^{t} \frac{i}{t})$ more computations compared with the first task. Apparently, it slowly grows with the total of tasks at a decreasing and negligible rate, resulting in an average growth of 1.04x and an average cost of 1.42x.

Self-supervised Iteration Analysis. Fig. 6c illustrates the performance of our method when adopting different self-supervised iterations (*i.e.*, e_{SSL} in Sec. 3.3). On this basis, we have the following observation:

(1) A too-small e_{SSL} (*e.g.*, 5) can cause degradation to the model, and results in even inferior performance compared with the baseline. We conjecture this is because too few SSL iterations may produce a poorly trained model.

(2) Appropriately increasing e_{SSL} leads to improved performance, and our method achieves the optimum with an acceptable cost, *i.e.*, $e_{SSL} = 15$.

(3) A too-large e_{SSL} also diminishes the performance, but still evidently surpasses the baseline. We deduce a plethora of SSL iterations may better maintain stability, but at the cost of compromising plasticity.

Additional Experiments. In supplementary material, we evaluate our approach with different memory sizes and demonstrate the generalization to other vision tasks.

5. Conclusion

In this work, we reformulate SSL, and provide an analytical solution to achieving instance-level discrimination, which addresses the trade-off widely embodied in current SSL strategies, hence yielding improved robustness against forgetting. On this basis, we design a novel alternate learning paradigm to enjoy the complementarity merits from both instance-level and category-level supervision, exhibiting significant superiority in maintaining stability and plasticity. Extensive experiments and sensitivities provide substantial evidence for the efficacy of our method.

ACKNOWLEDGMENT. This work is supported by the National Key Research and Development Program of China (2021ZD0111000), Science and Technology Commission (No.21511100700), National Natural Science Foundation of China (No.62222602, No.62106075, No.62176092), Natural Science Foundation of Shanghai (23ZR1420400), Shanghai Sailing Program (23YF1410500), Natural Science Foundation of Chongqing (CSTB2023NSCQ-JQX0007, CSTB2023NSCQ-MSX0137) and CAAI-Huawei Mind-Spore Open Fund.

References

[1] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9516–9525, October 2021.

[2] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*, 2019.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 2020.

[4] Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8250–8259, October 2021.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[6] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10164–10173, 2022.

[7] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision, (ECCV)*, pages 86–102. Springer, 2020.

[8] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9285–9295, 2022.

[9] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *Proceedings of the International Conference on Learning Representations*, 2020.

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning, ICML*, pages 1126–1135, 2017.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[12] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 831–839, 2019.

[13] Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, picking and growing for unforgetting continual learning. In *Advances in Neural Information Processing Systems*, pages 13647–13657, 2019.

[14] David Isele and Akansel Cosgun. Selective experience replay for lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[15] K J Joseph, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Vineeth N Balasubramanian. Energy-based latent aligner for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7452–7461, June 2022.

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[17] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[18] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations, (ICLR)*, 2020.

[19] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2544–2553, June 2021.

[20] Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations*, 2021.

[21] 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*, 2022.

[22] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

[23] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[24] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.

[25] Quang Pham, Chenghao Liu, and Steven Hoi. Dualnet: Continual learning, fast and slow. In *Advances in Neural Information Processing Systems*, 2021.

[26] Quang Pham, Chenghao Liu, Doyen Sahoo, and Steven HOI. Contextual transformation networks for online continual learning. In *International Conference on Learning Representations*, 2021.

[27] Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. itaml: An incremental task-agnostic meta-learning approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13588–13597, 2020.

[28] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2001–2010, 2017.

[29] 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 *International Conference on Learning Representations*, 2019.

[30] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence, (AAAI)*, volume 35, pages 9630–9638, 2021.

[31] Qing Sun, Fan Lyu, Fanhua Shang, Wei Feng, and Liang Wan. Exploring example influence in continual learning. In *Advances in Neural Information Processing Systems*, 2022.

[32] Shixiang Tang, Dapeng Chen, Jinguo Zhu, Shijie Yu, and Wanli Ouyang. Layerwise optimization by gradient decomposition for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9634–9643, June 2021.

[33] Yu-Ming Tang, Yi-Xing Peng, and Wei-Shi Zheng. Learning to imagine: Diversify memory for incremental learning using unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9549–9558, 2022.

[34] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations, ICLR*, 2019.

[35] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. FOSTER: feature boosting and compression for class-incremental learning. In *European Conference on Computer Vision (ECCV)*, 2022.

[36] Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 184–193, June 2021.

[37] Zhen Wang, Liu Liu, Yajing Kong, Jiaxian Guo, and Dacheng Tao. Online continual learning with contrastive vision transformer. In *European Conference on Computer Vision, (ECCV)*, 2022.

[38] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[39] Guile Wu, Shaogang Gong, and Pan Li Queen. Striking a balance between stability and plasticity for class-incremental learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1104–1113, 2021.

[40] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3014–3023, June 2021.

[41] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning, (ICML)*, pages 12310–12320, 2021.

[42] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5871–5880, June 2021.

[43] Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zhengjun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9286–9295, 2022.

[44] Chengxu Zhuang, Violet Xiang, Yoon Bai, Xiaoxuan Jia, Nicholas Turk-Browne, Kenneth Norman, James J. DiCarlo, and Daniel LK Yamins. How well do unsupervised learning algorithms model human real-time and life-long learning? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.