Knowledge-Based Systems 279 (2023) 110959

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

LLEDA—Lifelong Self-Supervised Domain Adaptation

Mamatha Thota [a], Dewei Yi [b], Georgios Leontidis [b,*]

[a] School of Computer Science, University of Lincoln, LN6 7TS, Lincoln, United Kingdom
[b] School of Natural and Computing Sciences & Interdisciplinary Centre for Data and AI, University of Aberdeen, AB24 3UE, Aberdeen, United Kingdom

ARTICLE INFO

Article history:
Received 19 February 2023
Received in revised form 4 August 2023
Accepted 30 August 2023
Available online 9 September 2023

Keywords:
Self-supervised learning
Representation learning
Life-long learning
Domain adaptation
Complementary learning systems
Latent replay

ABSTRACT

Humans and animals have the ability to continuously learn new information over their lifetime without losing previously acquired knowledge. However, artificial neural networks struggle with this due to new information conflicting with old knowledge, resulting in catastrophic forgetting. The complementary learning systems (CLS) theory (McClelland and McNaughton, 1995; Kumaran et al. 2016) suggests that the interplay between hippocampus and neocortex systems enables long-term and efficient learning in the mammalian brain, with memory replay facilitating the interaction between these two systems to reduce forgetting. The proposed Lifelong Self-Supervised Domain Adaptation (LLEDA) framework draws inspiration from the CLS theory and mimics the interaction between two networks: a DA network inspired by the hippocampus that quickly adjusts to changes in data distribution and an SSL network inspired by the neocortex that gradually learns domain-agnostic general representations. LLEDA's latent replay technique facilitates communication between these two networks by reactivating and replaying the past memory latent representations to stabilize long-term generalization and retention without interfering with the previously learned information. Extensive experiments demonstrate that the proposed method outperforms several other methods resulting in a long-term adaptation while being less prone to catastrophic forgetting when transferred to new domains.

1. Introduction

Deep neural networks have shown near human-level capabilities in many fundamental computer vision tasks [1–5]. Humans and animals can continuously acquire new information over their lifetime without catastrophically forgetting the prior knowledge learned. This ability to continually learn over time by accommodating new knowledge while retaining the previously learned knowledge is referred to as lifelong or continual learning (in our paper, we will continue to refer to it as lifelong learning). However, artificial neural networks lack these capabilities as new information interferes with previously learned knowledge and sometimes the old knowledge completely gets overwritten by the new one, leading to impaired performance [6]. The root cause of catastrophic forgetting is that learning necessitates changes in the weights of a neural network, however, these changes also result in the forgetting of previous learning.

The focus of this paper is on lifelong domain adaptation, in which the model is trained on multiple sequential domains, continuously adapting to new domains with changing distributions as they become available, while maintaining its knowledge of previously encountered domains.

Domain adaptation (DA) methods based on deep learning have received significant attention in recent years for mitigating the domain shift from the training domain to the inference domain [7–10], and have even been suggested as transformative technologies in settings such as agriculture [11,12] and arts [13]. However, current domain adaptation methods operate under the assumption that datasets from both the source and the target domains are accessible at the same time during training, which may not be feasible in practice. In addition, DA algorithms require fully labeled datasets, even state-of-the-art Unsupervised Domain Adaptation (UDA) methods need access at least to the source labeled dataset. Therefore, these algorithms require persistent manual annotation, which is time-consuming, cumbersome and expensive. Finally, just updating the underlying model will not be sufficient, as the model would likely forget the past learned domain information resulting in catastrophic forgetting. Acknowledging these issues, we propose LLEDA that addresses both catastrophic forgetting and domain-agnostic knowledge transfer using solely unlabeled datasets with access to a single domain at any given time.

The mammalian brain can continually acquire, process, consolidate, retrieve, and infer knowledge over time without catastrophically forgetting the previously learned information which can be explained using CLS theory [14,15]. It suggests that efficient learning in the mammalian brain requires two learning

* Corresponding author.
E-mail address: georgios.leontidis@abdn.ac.uk (G. Leontidis).

https://doi.org/10.1016/j.knosys.2023.110959

systems: the neocortex and the hippocampus. The first system gradually acquires structured generalized knowledge, while the second system quickly learns the specific experiences, and the interplay between these two systems enables long-term retention. It also implies that memory replay is the mechanism that facilitates interaction between these two systems to consolidate and stabilize new memories for long-term generalization to reduce catastrophic forgetting.

Recently, a study by [16] identified that the existing lifelong learning techniques are missing few biological elements. They highlight that many existing approaches solely focus on modeling the cortex directly and do not have a rapid learning network which is essential for facilitating effective lifelong learning in the brain. Additionally, the study also points out that none of the current methods employs information from the neocortex-inspired network to influence the training of the hippocampal-inspired network, whereas, in biological networks, the neocortex influences learning in the hippocampus and vice versa.

Our proposed LLEDA network attempts to solve the first issue by utilizing two distinct networks, DA network for rapid learning and the SSL network for gradual acquisition. LLEDA mimics the interplay between the neocortex and the hippocampus, where the hippocampal-inspired DA network functions as a rapid acquisition mechanism to adapt the distribution shift between the given data stream and the data from memory, and the neocortex-inspired SSL network works like a gradual learning mechanism to generalize the representations by gradually acquiring structured knowledge using self-supervised techniques enabling effective lifelong learning. LLEDA's Latent memory replay facilitates communication between these two networks by reactivating the neural activity patterns representing previous experiences to stabilize new memories for long-term generalization and retention without interfering with the previously learned information. LLEDA attempts to address the second issue by querying the information from the neocortex to influence the training of the hippocampal-inspired network during training.

Overall, our framework reduces catastrophic forgetting, while facilitating domain-agnostic knowledge transfer without accessing labeled data both from the source and target domains at any given time. To the best of our knowledge, this is an area of domain adaptation that has not yet been explored. In summary, our work makes the following contributions:

1. Inspired by the CLS theory, LLEDA mimics the interplay between the DA network which helps to rapidly adapt the distribution shifts between domains, and the SSL network that helps with the gradual acquisition of domain-agnostic general representations, and the latent representations replay technique helps to replay the past memory representations, instead of raw image pixels to overcome catastrophic forgetting.
2. Our proposed self-supervised based approach does not require access to either source or target labels, hence saving time and effort to annotate data and assisting with the labeling bias.
3. Extensive empirical results demonstrate that our method performs competitively across several benchmarks, when compared against other approaches.

The rest of the paper is organized into several sections. Section 2 offers an extensive literature review on Domain Adaptation (DA), Self-Supervised Learning (SSL), and Continual Learning (CL). In the Section 3, we present a detailed explanation of our LLEDA framework, which comprises three main components: Generalized Feature Learning, Domain-Specific Representation Learning, and Latent Replay. The Section 4 discusses the datasets used, training methodology, implementation details, and presents results, analysis, and ablation studies. Finally, Section 5 provides a summary of the paper's findings and outlines potential directions for future research.

2. Related work

Domain Adaptation: Under the assumption of independent and identically distributed (iid) data, a deep neural network trained on one set of data is expected to perform well on a new, unseen set of data. However, this assumption may not always hold in real-world applications due to the discrepancy between domain distributions, and applying the trained model to the new dataset may also result in negative performance. Domain adaptation is a special case of transfer learning where the goal is to learn a discriminative model in the presence of domain shift between source and target datasets. Various methods have been introduced to minimize the domain discrepancy in order to learn domain-invariant features. Some involve adversarial methods like DANN [9], ADDA [17] that help align source and target distributions. Other methods propose aligning distributions through minimizing divergence using popular methods like maximum mean discrepancy [3,7,8,10,18–21], correlation alignment [22–24], and the Wasserstein metric [25,26]. MMD was first introduced for the two-sample tests of the hypothesis that two distributions are equally based on observed samples from the two distributions [18], and this is currently the most widely used metric to measure the distance between two feature distributions. The Deep Domain Confusion Network proposed by Tzeng et al. [27] learns both semantically meaningful and domain invariant representations, while Long et al. proposed DAN [7] and JAN [19] which both perform domain matching via multi-kernel MMD (MK-MMD) or a joint MMD (J-MMD) criteria in multiple domain-specific layers across domains.

Self-Supervised Learning: Self-Supervised Learning (SSL) is a paradigm developed to learn visual features from unlabeled data. Recently, numerous SSL approaches have shown significant performance sometimes even surpassing, the performance of supervised baselines [28–39]. These methods use image augmentation techniques to generate multiple views of a given image and learn a model that is invariant to these augmentations. Most recent approaches are divided into two main categories, contrastive and non-contrastive methods. Contrastive methods learn an embedding space where positive pairs are pulled together, whilst negative pairs are pushed away from each other [28–30]. Non-contrastive methods on the other hand remove the need for explicit negative pairs either by using distillation or by regularization of the variance and covariance of the embeddings [32–35]. However, none of these works studied the ability of SSL methods to learn continually and adaptively if they are applied directly. Moreover, very few works have attempted to use SSL in the lifelong domain adaptation setting, e.g. [40] is designed using contrastive learning, so it lacks the capability to adapt using other SSL paradigms. [41] trains model stepwise by generating pseudo labels and fine-tuning on intermediate domains until it reaches the target domain, this model can adapt well only if the domain shift is small between the intermediate domains, and it also uses source-labeled data. In this paper, we present a general-purpose framework to incorporate self-supervised learning approaches into the lifelong learning process to extract generalized representations.

Continual learning: Continual learning strategies aim to find the right balance between preventing catastrophic forgetting and acquiring new information. According to [42], catastrophic forgetting can be mitigated using model regularization, memory replay or by expanding and training the network. Regularization

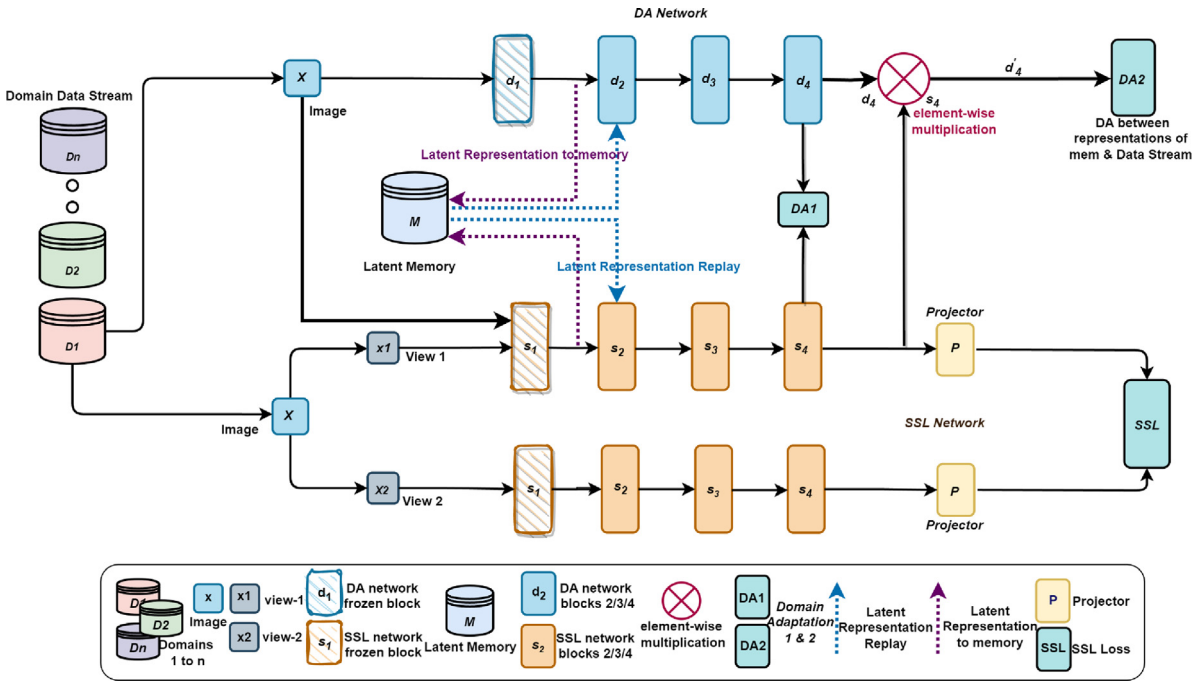

Fig. 1. Overview of the proposed LLEDA architecture. LLEDA consists of rapid learning **DA network** and gradual learning **SSL network**. The **SSL network** learns generic representations using self-supervised learning and **DA network** helps to overcome domain shift by optimizing DA loss at two levels, **DA1**- MMD loss between the representations of d_4 and s_4, and **DA2** - MMD loss between memory representations and current data representations.

methods identify the network weights that contribute significantly to retaining knowledge about a previously learned task and then consolidate them when the model is updated to learn the subsequent tasks [43–45]. On the other hand, dynamic architectures modify the model's underlying architecture by dynamically accommodating neural resources as it learns new patterns [46–48]. Alternatively, the model can be expanded progressively to learn the new tasks using added weights that propose ways of constraining the tasks' objectives to avoid forgetting [49–51]. CLS and replay methods rely on memory replay by storing samples from old distributions and regularly feeding them back to the model to overcome catastrophic forgetting. Some of the existing CL methods [52–54] store raw inputs of previous data in the memory, however, replaying raw pixels is not biologically plausible. Generative replay methods involve training a generative model like an auto-encoder or a generative adversarial network to produce samples from previously learned data [55,56]. However, these approaches are very difficult to train due to issues such as convergence and mode collapse, additionally scaling up generative replay to complex datasets is challenging. Latent replay methods involve storing compressed representations at a specific layer, rather than keeping duplicate copies of input patterns as raw data. These compressed representations capture the essential features of the input data, making them efficient for replay. Utilizing latent replay in LLEDA is not only the most efficient but also a biologically plausible approach [57–59]. We summarize LLEDA as follows:

- Existing research on combining the lifelong learning and domain adaptation is limited. While some studies like [60] focus on continual and supervised adaptation using labeled data, others such as [61,62] address continual domain adaptation but assume gradual target shifts, making them less practical.
- We present a novel solution called LLEDA, which draws inspiration from the mammalian brain and the CLS theory. LLEDA addresses the issue of catastrophic forgetting and facilitates domain-agnostic knowledge transfer, operating exclusively with unlabeled datasets, allowing learning from a single domain at a time.
- LLEDA lies at the intersection of lifelong learning, self-supervised learning, and domain adaptation.

3. Methodology

Our overall objective is to continually update a model to learn distributional shifts while retaining knowledge about past learnings. We propose a novel lifelong domain adaptation framework (depicted in Fig. 1 and algorithm 1), which has three key components and is motivated by the CLS theory [14]. The DA network in LLEDA swiftly adapts to changes in the data distribution between the current domain and previously encountered domains. The SSL network learns to generalize representations through self-supervised learning of domain-agnostic data, while the latent memory component facilitates the interaction between the two networks. By replaying and reactivating past experiences, this component stabilizes new memories for long-term retention and generalization. The combined operation of the DA and SSL networks integrates new information into the long-term network without compromising previous knowledge.

The LLEDA framework process involves the following steps: first, the SSL network learns the visual features and their relationships from the unlabeled input data using self-supervised techniques. As the SSL network is not task-specific, the learned representations are more general, capturing the underlying structure of the data. Next, the DA network uses Maximum Mean Discrepancy (MMD) loss to address domain shift between the current domain and previous domains stored in memory. This loss is backpropagated to both networks for consolidation and to prevent interference. The latent memory component stores and replays past experiences as representations, rather than raw input pixels, to aid interaction between the two networks. All learning occurs in a synchronous and interleaved manner.

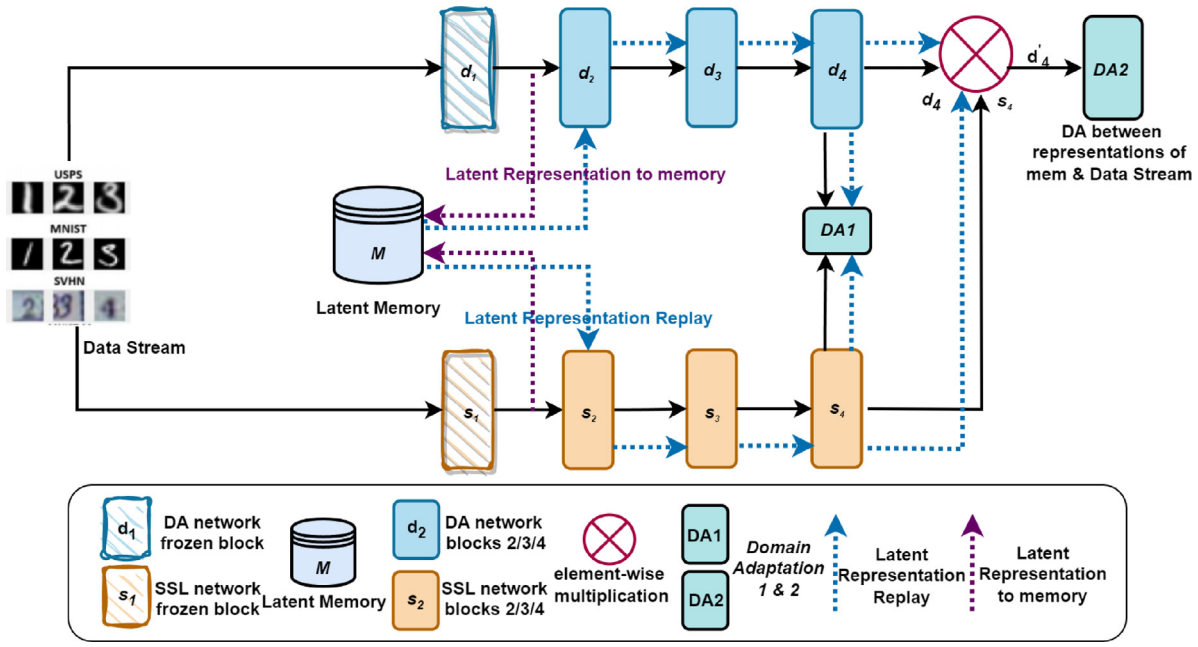

Fig. 2. Overview of latent replay. Demonstration of the flow of latent representations, **the arrows in blue** show the latent representation flow from memory to the network and **arrows in pink** show the flow of latent representations from network to memory.

3.1. Generalized feature learning

LLEDA employs the SSL network to gradually learn and capture the visual features, underlying structure, and their relationship. As this network is trained independently from the DA network, it does not interfere with the new learnings. Moreover, self-supervised learning provides the model with additional context and information about the input data, enabling it to learn more generic and transferable representations in situations where labeled data is not accessible, which is the case in our scenario.

LLEDA's SSL backbone network is compatible with all the existing SSL models (SimCLR [28], BYOL [34], etc.,), so any generic SSL model can be used as the backbone. However, we have considered VICReg [63] as our backbone to reduce the SSL loss due to its simplicity, additionally it does not require a memory bank, contrastive samples, or a large batch size. We have conducted ablation studies using alternative SSL models like SimCLR [28] and BYOL [34] as our background network to reduce the SSL network's loss, which has been discussed later in Section 4.5. VICReg model uses the weighted average of invariance, variance and covariance to calculate the loss between . The SSL loss is defined as follows:

$$l(z_i, z_j) = \lambda s(z_i, z_j) + \mu[\upsilon(z_i) + \upsilon(z_j)] + \nu[c(z_i) + c(z_j)] \tag{1}$$

Where λ, μ, ν are the hyper-parameters controlling the importance of each term in the loss. $s(z_i, z_j)$ is the Invariance, $c(z_i)$, $c(z_j)$ is covariance and $\upsilon(z_i)$, $\upsilon(z_j)$ is variance.

The overall objective is given by

$$L = \sum_{I \epsilon D} \sum_{t_i, t_j \sim T} l(z_i, z_j) \tag{2}$$

3.2. Domain-specific representations learning

The goal of the DA network is to rapidly learn to reduce the domain discrepancy for the incoming domains, simultaneously working well on the previous domains without catastrophically forgetting the learnings. The DA network uses Maximum Mean Discrepancy (MMD) loss to address domain shift. The DA network, inspired by Dualnet [54], also interacts with the SSL network and acquires generic representations that influence its learning in a manner akin to biological networks, improving its capacity to reduce discrepancy between domains. It reduces the discrepancy in two stages: The DA network uses Maximum Mean Discrepancy (MMD) loss to address domain shift. It calculates MMD loss using representations from block 4 of the Resnet (DA1). It again calculates the MMD loss between the memory representations and the current data stream propagation (DA2) following the element-wise multiplication. Calculating the MMD loss at two stages (DA1 and DA2), as seen in Fig. 2, helps to effectively reduce the domain shift, compared to a single domain adaptation loss.

Let s_4 be the feature representation from the SSL network's residual block, and d_4 be the feature representation from the DA network's residual block as shown in Fig. 2, the adapted feature is obtained during network interaction as follows:

$$\mathrm{d}_4' = \mathrm{d}_4 \otimes \mathrm{s}_4 \tag{3}$$

where $\otimes$ denotes the element-wise multiplication, the output of the rapid DA network d_4, gradual SSL network s_4 and the transformed feature $d_4{}'$ all have the same dimension.

The final layer's transformed feature $d_4{}'$ will be fed into the DA network's head to calculate the DA2 loss using MMD. The rapid DA network takes advantage of the gradual SSL learner's generalized feature representations resulting in quick adaptation leading to reduced domain shift and improved generalization leading to better identification of classes in the downstream classification task.

MMD defines the distance between the two distributions with their mean embeddings in the Reproducing Kernel Hilbert Space (RKHS). MMD is a two-sample kernel test to determine whether to accept or reject the null hypothesis $p = q$ [18], where p and q are source and target domain probability distributions. In short, the MMD between the distributions of two datasets is equivalent to the distance between the sample means in a high-dimensional feature space and is computed by the following equation:

$$L_{MMD} = \left\| \frac{1}{N} \sum_{i=1}^{N} \phi(x_i^s) - \frac{1}{M} \sum_{j=1}^{M} \phi(x_j^t) \right\|_H^2 \tag{4}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N} k(x_i^s, x_{i'}^s) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} k(x_i^s, x_j^t) + \frac{1}{M^2}\sum_{j=1}^{M}\sum_{j'=1}^{M} k(x_j^t, x_{j'}^t) \quad (5)$$

where: $\phi(.)$ is the mapping to the RKHS H; and $k(.,.) = \langle \phi(.), \phi(.) \rangle$ is the universal kernel associated with this mapping, and N, M are the total number of items in the source and target respectively.

3.3. Latent replay

The mammalian brain has successfully evolved to resist catastrophic forgetting by reactivating, replaying, and recreating the experience preserved in memories [64,65]. It retains compressed versions of the crucial information from past experiences and reactivates by replaying these neural activity patterns of prior experiences. Inspired by this, LLEDA stores feature representations from a specific layer instead of raw input pixels. By doing so, it reactivates and replays these representations to overcome catastrophic forgetting. To achieve this, we freeze the layers below the chosen layer, effectively preventing them from being updated during training.

We implement parameter freezing by freezing the layers below block-1, which effectively disables gradient updates for these layers. As a result of this process, the weights in the frozen layers remain unchanged, while the weights in the unfrozen (trainable) layers are allowed to adapt based on the new data. This strategy ensures the preservation of integrity for these layers, promoting stability and accuracy during the replay of stored representations, while also mitigating any potential aging effect [57]. Freezing the network also helps with the stability of the stored representations, else they will differ from the feature representations that would have been generated while feed-forwarding from the input layer. In LLEDA, we save the representations from block-1 of our backbone ResNet network into memory and freeze the network layers below block-1 (below the latent replay layer) to prevent them from being updated during the subsequent training on a new task or dataset and to ensure the stability and accuracy of the representations and to prevent the aging effect.

As our model does not have access to labels, we follow a simple approach of storing a random subset of past latent representations in memory and train the network while interleaving with new domain representations [66]. While selective replay has shown promising results in few settings, several studies have found that random sampling works equally well [58,67], achieving similar performance making it a computationally efficient choice, hence we store random subset of representations in the memory. Following that, we save the latent representations from both the DA and the self-supervised networks for the given random image. During memory consolidation, these memories are interleaved with new latent representations to form a more general representation supporting long-term retention and generalization when encountering new domain experiences. To avoid inefficiency, we store only a limited number of latent representations per domain in the memory buffer until it reaches a given number, known as the latent memory size. In our experiments, we tested two sizes: 100 and 250 latent representations. This ensures that the buffer contains a manageable amount of past random experiences at any given time, as depicted in algorithm 2.

Algorithm 1: Pseudocode for the proposed Lifelong Domain Adaptation

Input : Current Domain Data D, Memory M, SSL θ, DA ϕ
Output: updated θ, ϕ
for *sampled minibatch (S_d, S_m) from D and M* **do**
  Calculate L_{SSL} loss on S_d using equation: (2) to update θ
  Calculate L_{DA1} loss on S_d using equation: (5) to update ϕ and θ
  if *domain* > 1 **then**
    Calculate L_{DA1} loss on S_m using equation: (5) to update ϕ and θ
    Calculate L_{DA2} loss on S_d and S_m using equation: (5) to update ϕ and θ
  end if
  Add latent representations to memory using algorithm:2
end for

Algorithm 2: Pseudocode for saving random latent representations to memory

Input : Memory M, Representations R, Sample Size s
Output: Memory M
$M = \theta$
t_m = len(M)
c_m = 0
for *each repbatch from R* **do**
  $\delta = t_m - c_m$
  $h = min(s, \delta)$
  R_{add} = random sampling of size h from repbatch
  if $c_m < t_m$ **then**
    $M = M \cup R_{add}$
    $c_m += h$
  else
    $R_{replace}$ = random sampling of size s from M
    $M = (M - R_{replace}) \cup R_{add}$
  end if
end for

4. Experiments & results

4.1. Datasets

We compare and evaluate our method against baseline approaches on a number of benchmark domain adaptation datasets, such as Digits, Office-Home [68], Office-CalTech [69] and ImageCLEF-DA.

Digit Dataset: We consider the standard digits dataset broadly adopted by the computer vision community. MNIST [70] and USPS [71] are hand-written grey-scale images, with relatively small domain differences. SVHN [72] contains images of street numbers with more than one digit in each image. We conducted experiments on two tasks: SVHN → USPS → MNIST and MNIST → USPS → SVHN and reported the average accuracy of the trained model in the context of lifelong learning setting. These two scenarios will allow us to reflect on the performance of lifelong learning scenarios starting from easy datasets, moving to harder ones and vice versa. Sample images of the digit dataset are presented in Fig. 3.

Office-Home [68]: The office-home data consists of four visual domains: Art (A), Clipart (C), Real World (R), and Product (P) each consisting of images from 65 visual categories totaling 15,500 images in office and home settings leading to the possibility of defining 12 pair-wise binary UDA tasks. We conducted several

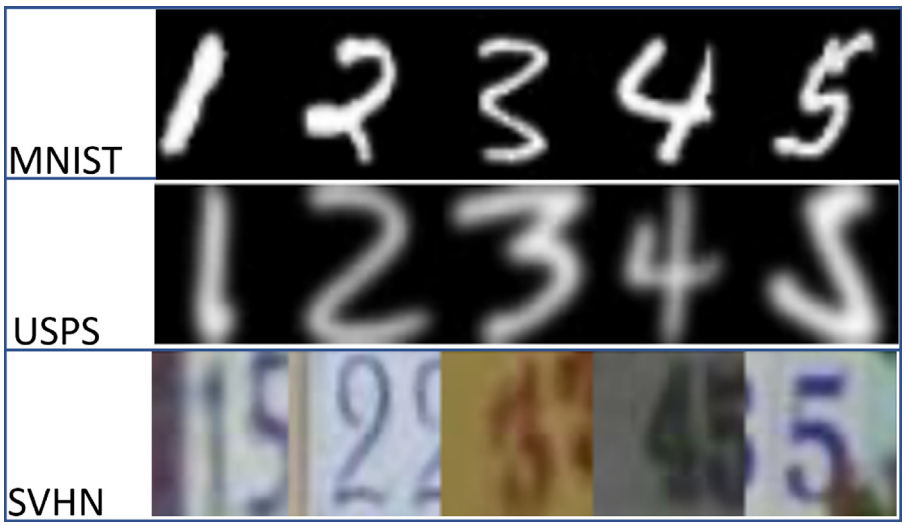

Fig. 3. Sample images from digits dataset.

Fig. 4. Sample images from office-home dataset.

experiments on two tasks: Art → Realworld → Clipart → Product and Product → Clipart → Realworld → Art and reported the average accuracy of the trained model in the context of lifelong learning setting. Sample images of the office-home dataset are presented in Fig. 4.

Office-CalTech [69]: This dataset is an extension of the Office-31 [73] with 10 common categories shared by Office-31 and the CalTech-256 dataset [74]. This dataset has four domains: Webcam (W), DSLR (D), Amazon (A), and CalTech (C). We conducted several experiments on two tasks: DSLR → Webcam → Amazon → Caltech and Caltech → Amazon → Webcam → DSLR and reported the average accuracy of the trained model in the context of lifelong learning setting. Sample images of the office-caltech datasets are presented in Fig. 5.

ImageCLEF-DA: This dataset has four domains with twelve categories each: Caltech-256, ImageNet ILSVRC 2012, and Pascal VOC 2012. We conducted several experiments on two tasks: Caltech → ImageNet → Pascal and Pascal → ImageNet → Caltech and reported the average accuracy of the trained model in the context of lifelong learning setting. Sample images of the ImageCLEF dataset are presented in Fig. 6.

4.2. Training methods

We benchmark LLEDA against the baseline method which uses a single network and finetunes the model as the new training domains come along, we then compare our LLEDA methodology with DANN [9] and DAN [7], both of them are classic domain adaptation methods and both these methods have access to source and target data during training. We also compare LLEDA with CUA [62] and GRCL [40] which are continual learning replay-based methods. It is important to note that both methods have access to source-labeled data, unlike LLEDA which operates without labeled data from either the source or target domains. We made an exciting observation during experimentation by increasing the size of latent representations stored in memory from 100 to 250, resulting in impressive results. Consequently, we tested LLEDA-100 and LLEDA-250, with 100 and 250 latent

Fig. 5. Sample images from office-caltech dataset.

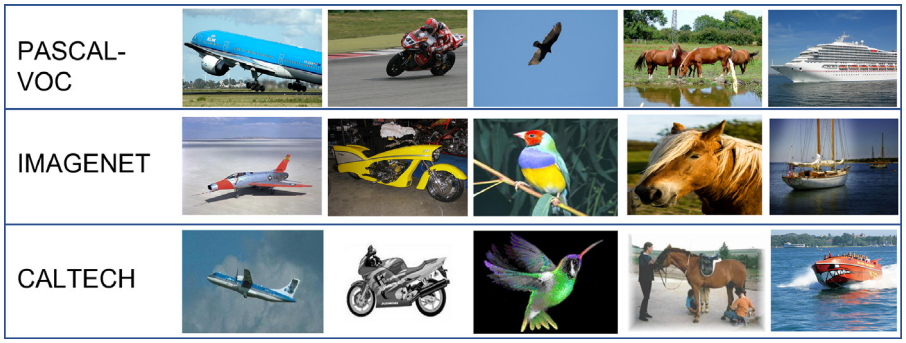

Fig. 6. Sample images from ImageCLEF-DA dataset.

representations stored in the memory respectively. It is worth mentioning that other papers typically use a memory size of 2000 images/representations. Furthermore, we compare these methods with the supervised version of our approach, LLEDA-S. Most of the methods provide the results in the domain adaptation setting, but we have provided our results in the context of lifelong learning setting.

4.3. Implementation details

Our implementation involves three stages. In the first stage, we pre-train the model on ImageNet which serves as the foundation for subsequent stages, which we call as a pre-trained model. In the second stage, we use the pre-trained model and further train the LLEDA model as outlined in the methodology section. In the final stage, we freeze the trained network and train a linear classifier on top of the fixed representation, while removing the MMD projection head. This trained linear classifier is used for evaluation purposes.

For the pretraining phase, we employ the ResNet18 [75] architecture as our backbone model, pretrained on the ImageNet dataset. During this phase, we use two nodes, each equipped with 4 V100 GPUs. The training process is carried out using the LARS optimizer [76] with a batch size of 512, and we apply a weight decay of le-6 training for a total of 100 epochs. During the subsequent training phase, we use the pretrained network obtained from the previous stage as a starting point. We incorporate the stored latent representations from the layer-1 and combine it with the current domain data representations. Finally, during finetuning phase, we freeze the trained network and further train a linear classifier on top of this fixed representations whilst discarding the MMD part of the network. We use the resulting network to evaluate on the domain datasets to assess its performance. Similar to most self-supervised models [28,29,29,30,32–35,77], we report performance by training a linear classifier on top of a fixed representation to evaluate representations which is a standard benchmark that has been adopted by many papers in the literature.

4.4. Results and analysis

Our primary objective is to evaluate the performance of our proposed LLEDA framework in lifelong learning domain adaptation scenarios. This assessment involves sequentially training the model on different domains. This sequential training process, which we refer to as a "cycle", involves training the model on one domain, followed by training it on the next domain, and so on. Upon the completion of each cycle, we consider the resulting model as the "final model". This final model is then tested on all the domains it was trained on, and the corresponding results are presented. Additionally, we calculate the average performance

Table 1
Comparison of the proposed LLEDA method on Digit datasets comprising MNIST, USPS and SVHN domains with state-of-the-art methods, using Average Accuracy (Avg) across the domains as the performance metric. LLEDA-100 and 250 represent the latent memory size of 100 and 250. LLEDA-S is a supervised model with access to labels. The best average is indicated in **bold**.

Dataset	Method	Avg
Digits	Baseline	56.7
	DANN	74.5
	DAN	72.9
	CUA	82.1
	GRCL	85.3
	LLEDA-S	89.0
	LLEDA-100	86.6
	LEDA-250	**89.5**

Table 2
Comparison of the proposed LLEDA method on Office-Home datasets comprising Art, Clipart, Product and Real-World datasets with state-of-the-art methods, using Average Accuracy (Avg) across the domains as the performance metric. LLEDA-100 and 250 represent the latent memory size of 100 and 250. LLEDA-S is a supervised model with access to labels. The best average is indicated in **bold**.

Dataset	Method	Average
Office-Home	Baseline	28.7
	DANN	57.6
	DAN	56.3
	CUA	58.6
	LLEDA-S	60.3
	LLEDA-100	58.2
	LLEDA-250	**62.1**

of the final model across each domain within every cycle. It is crucial to note that the other state-of-the-art methods, which we refer to, operate based on the UDA evaluation criteria. These methods are trained on a labeled source dataset and tested on an unlabeled target dataset, but they are not continually trained as in our case. Therefore, our results reflect the challenging scenario where testing is performed within cycles, and without access to any labeled data.

Baseline: Initially, we train a basic model, denoted as M_i, on the domain D_i. As new domains become available, we fine-tune the model by training it on the subsequent domain, D_{i+1}. However, we observe that at the end of the cycle, this approach tends to exhibit poor performance on earlier domains due to a phenomenon known as CF. This outcome serves as our baseline for comparison in this study. In our experiments, we use Resnet18 as the baseline model and to assess the effectiveness of our proposed method, we evaluate its performance against the baseline.

Digits dataset: The Table 1 presented in this study showcases the average performance of different methods on the Digits dataset, encompassing MNIST, USPS, and SVHN domains in the lifelong learning scenario cycle. The proposed method, LLEDA, exhibits a remarkable advantage over other approaches by effectively handling sequential training in lifelong learning adaptation scenarios. LLEDA achieves an impressive average accuracy of 89.5%, surpassing the baseline accuracy of 56.7%, and outperforms several state-of-the-art methods, including DANN (74.5%), DAN (72.9%), CUA (82.1%), and GRCL (85.3%). These results underscore the superior performance of LLEDA in the context of lifelong learning, highlighting its adaptability and sequential learning capabilities. It is worth noting that LLEDA-S demonstrates improved performance compared to LLEDA-100, which can be attributed to the availability of labeled data. Additionally, the performance gap between LLEDA-100 and LLEDA-250 can be linked to the amount of additional memory representations saved. With more representations stored and replayed, LLEDA-250 accumulates a diverse set of samples from various domains or tasks, enabling the model to develop more robust and generalizable features, resulting in enhanced performance on new data.

Office-Home dataset: The Table 2 presents the average performance of different methods on the Office-Home dataset involving Art, Clipart, Product and Real-world domains, and focuses on the lifelong domain adaptation cycle. The baseline method achieves an average accuracy of 28.7%, which is relatively low. In contrast, our proposed LLEDA method without access to labels, achieves a substantial increase in average accuracy, rising from 28.7% to an impressive 62.1%. Similar to the findings in the digits dataset, LLEDA-S outperforms LLEDA-100 as expected. Moreover, LLEDA-250 achieves the highest average accuracy among the other state-of-the-art methods suggesting its superior performance effectively handling the lifelong learning adaptation scenarios on the Office-Home dataset.

Office-Caltech dataset: The Table 3 presents the average accuracy of various methods on the Office-Caltech dataset, along with Amazon, Caltech, DSLR and Webcam domains. It summarizes the average performance of different methods on the Office-Caltech dataset. The baseline method achieves an average accuracy of 52.3%, while DANN performs significantly better at 81.7%. EWC and CUA further improve the results with average accuracies of 84.5% and 84.8%, respectively. GRCL shows even higher performance with 87.2% accuracy, and LLEDA-S follows closely with 87.5% accuracy. LLEDA-100 achieves an average accuracy of 86.1%, and the top-performing method in this dataset is LLEDA-250, achieving an impressive average accuracy of 90.3% without utilizing any labeled data. These results demonstrate the superiority of LLEDA in effectively addressing lifelong learning adaptation scenarios within the Office-Caltech dataset, outperforming other state-of-the-art methods and indicating its potential as a robust approach for continual domain adaptation tasks.

ImageCLEF-DA dataset: The Table 4 presents the average accuracy of various methods on the ImageCLEF dataset, along with Pascal VOC, ImageNet, and Caltech domains. It summarizes the average performance of different methods on the ImageCLEF-DA dataset. The results indicate that the ImageCLEF Baseline method achieved the lowest average accuracy of 51.3%, suggesting limited performance in handling domain shifts. DANN and DAN, which employ domain adaptation techniques, showed significant improvement with average accuracies of 82.2% and 82.4%, respectively. However, their performance was surpassed by the LLEDA-250 method, which achieved an impressive average accuracy of 92.6%. Both LLEDA-100 and LLEDA-250 outperformed DANN and DAN, with average accuracies of 89.7% and 92.6%, respectively. Similar to the domains above LLEDA-S has an increased performance in comparison to LLEDA-100 due to the access of labeled data. LLEDA-250 emerged as the top-performing method across all datasets, suggesting its superiority in addressing continual domain adaptation tasks effectively.

The experimental results conducted on the Digits, Office-Home, Office-Caltech and ImageCLEF-DA datasets consistently demonstrate the superior performance of the LLEDA framework over other state-of-the-art methods in addressing lifelong learning scenarios. LLEDA effectively tackles challenges such as catastrophic forgetting and domain shift through sequential access to unlabeled data, showcasing its adaptability. Despite variations in evaluation setups, LLEDA showcases impressive performance in handling domain shifts even without label access, resulting in consistently high average accuracy. These results affirm the robustness and potential of the LLEDA framework in addressing lifelong learning challenges across diverse datasets.

Table 3
Comparison of the proposed LLEDA method on Office-Caltech datasets comprising Amazon, Caltech, DSLR and webcam domains with state-of-the-art methods, using Average Accuracy (Avg) across the domains as the performance metric. LLEDA-100 and 250 represent the latent memory size of 100 and 250. LLEDA-S is a supervised model with access to labels. The best average is indicated in **bold**.

Dataset	Method	Average
Office-Caltech	Baseline	52.3
	DANN	81.7
	EWC	84.5
	CUA	84.8
	GRCL	87.2
	LLEDA-S	87.5
	LLEDA-100	86.1
	LLEDA-250	**90.3**

Table 4
Comparison of the proposed LLEDA method on ImageCLEF-DA datasets comprising Caltech, ImageNet ILSVRC, and Pascal-VOC domains with state-of-the-art methods, using Average Accuracy (Avg) across the domains as the performance metric. LLEDA-100 and 250 represent the latent memory size of 100 and 250. LLEDA-S is a supervised model with access to labels. The best average is indicated in **bold**.

Dataset	Method	Average
ImageCLEF	Baseline	51.3
	DANN	82.2
	DAN	82.4
	LLEDA-S	90.3
	LLEDA-100	89.7
	LLEDA-250	**92.6**

Table 5
Comparison of the proposed LLEDA's SSL network using state-of-the-art self-supervised methods as building blocks.

Method	CYCLE-1			CYCLE-2			Avg
	SVHN	USPS	MNIST	MNIST	USPS	SVHN	
LLEDA-VICReg	71.3	93.3	94.1	86.7	85.9	88.7	86.6
LLEDA-SimCLR	73.6	94.8	93.8	78.9	87.2	90.5	86.4
LLEDA-BYOL	70.9	95.5	92.6	86.3	88.9	87.5	86.9

Table 6
Comparison of LLEDA's DA and SSL network interaction using various element-wise operations. The best average is indicated in **bold**.

Dataset	Method	Average
Digits	Elementwise multiplication	**86.6**
	Elementwise addition	75.3
	Elementwise maximum	39.3
	Elementwise mean	71.9

4.5. Ablation studies

Ablation: LLEDA's SSL network using state-of-the-art self-supervised methods as building blocks We evaluated the effectiveness of LLEDA by replacing the LLEDA's SSL network with some of the state-of-the-art SSL networks. Our objective is to assess the LLEDA's performance in lifelong learning scenarios by sequentially training on different domains. To assess lifelong learning performance, we start by training the image samples from one domain, followed by training on the next domain, and so on. We call this sequential training process as a cycle. For example, in cycle-1 (SVHN - USPS - MNIST), we trained the LLEDA model on the SVHN, followed by training on the USPS, and finally on the MNIST. Similarly, in cycle-2 (MNIST - USPS - SVHN), we trained on MNIST followed by training on USPS, and finally training on SVHN. Each cycle represents a sequential training process on different datasets.

We analyzed the accuracy of LLEDA to investigate the impact of the SSL network selection on gradual learning network. In Table 5, we compare three SSL methods- SimCLR [28], BYOL [34] and VICReg. We chose these SSL networks as all three methods feature different losses and use different techniques to avoid collapse such as negative samples, redundancy reduction, etc. Additionally, the former is a contrastive-based method, whereas the latter two are non-contrastive ones.

Table 5 shows that the average performance of VICReg is robust in comparison to the average performance of contrastive-based SimCLR [28] as the latter requires large amounts of contrastive pairs and a higher batch size to converge. The average performance of VICReg slightly underperforms compared to BYOL [34]. Overall, the comparative performance of all three SSL methods with respect to the LLEDA framework is almost relatively similar, with minor variations in accuracy across different datasets and cycles. This similarity can be attributed to the fact that SSL is not directly employed for the downstream task in LLEDA. Instead, element-wise multiplication balances the contributions of both SSL and DA networks. This network integration into the downstream tasks may explain the comparable performance across various SSL techniques. Consequently, these findings suggest that LLEDA's gradual learning network can effectively accommodate the substitution of any generic SSL method, ensuring its efficiency and adaptability to different SSL approaches.

LLEDA SSL and DA network interaction using element-wise operations: We analyzed different types of operations used for interactions and influence between the DA network and the SSL network. We considered element-wise addition, element-wise maximum value, and element-wise mean besides adapting element-wise multiplication to test the generalization ability. Table 6 demonstrates that the element-wise maximum value seems like a poor choice since the interaction between the two networks appears more competitive than complementary. Element-wise multiplication excels at emphasizing agreement between networks, resulting in a combined representation that is more domain-invariant and robust. In contrast, element-wise addition and element-wise show some promise compared to element-wise maximum value, but fall short in capturing complementary features and striking an optimal balance between SSL and DA networks, ultimately leading to subpar model performance. Therefore, the adapted element-wise multiplication emerges as the most favorable choice offering superior generalization capabilities.

5. Conclusion & future work

Inspired by how the human brain works and the CLS theory, we developed LLEDA, a model that can perform competitively in a lifelong domain adaptation setting across several standard benchmark datasets. Our experiments demonstrate that LLEDA can effectively tackle downstream domain adaptation tasks without access to labeled data, outperforming several other existing methods. This is a very exciting line of work, as in many real-life settings, e.g. healthcare and agriculture, one does not have the luxury of large, curated and labeled data. With methods like LLEDA, we can leverage such datasets to learn continuously and improve performance.

We believe our work will encourage future research in lifelong domain adaptation using unlabeled source and target domain data, as this is a more realistic scenario in several real-life settings. As our next step, we aim to investigate efficient lossy and

lossless compression techniques for compressing latent representations in LLEDA, as well as show how LLEDA performs in larger datasets, which require extensive computational resources. Another line of work will involve exploring techniques, e.g. through distillation and quantization, that will reduce the computational overhead required.

CRediT authorship contribution statement

Mamatha Thota: Conceptualization, Data curation, Investigation, Methodology, Software, Formal Analysis, Writing – original draft. **Dewei Yi:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Methodology, Supervision, Writing – review & editing. **Georgios Leontidis:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

This work used the Cirrus UK National Tier-2 HPC Service at EPCC (http://www.cirrus.ac.uk). Access granted through the project: ec173 - Next-gen self-supervised learning systems for vision tasks.

References

[1] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[2] F. He, Y. Wang, X. Miao, X. Sun, Interpretable visual reasoning: A survey, Image Vis. Comput. 112 (2021) 104194.
[3] L. Gong, M. Thota, M. Yu, W. Duan, M. Swainson, X. Ye, S. Kollias, A novel unified deep neural networks methodology for use by date recognition in retail food package image, Signal Image Video Process. 15 (3) (2021) 449–457, http://dx.doi.org/10.1007/s11760-020-01764-7.
[4] F.D.S. Ribeiro, K. Duarte, M. Everett, G. Leontidis, M. Shah, Learning with capsules: A survey, 2022, arXiv preprint arXiv:2206.02664.
[5] H. Ren, W. Lu, Y. Xiao, X. Chang, X. Wang, Z. Dong, D. Fang, Graph convolutional networks in language and vision: A survey, Knowl.-Based Syst. (2022) 109250.
[6] M. McCloskey, N.J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of Learning and Motivation, Vol. 24, Elsevier, 1989, pp. 109–165.
[7] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, PMLR, 2015, pp. 97–105.
[8] M. Thota, S. Kollias, M. Swainson, G. Leontidis, Multi-source domain adaptation for quality control in retail food packaging, Comput. Ind. 123 (2020) 103293.
[9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (1) (2016) 2030–2096.
[10] M. Thota, G. Leontidis, Contrastive domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 2209–2218.
[11] A. Durrant, M. Markovic, D. Matthews, D. May, G. Leontidis, J. Enright, How might technology rise to the challenge of data sharing in agri-food? Glob. Food Secur. 28 (2021) 100493.
[12] G. Onoufriou, M. Hanheide, G. Leontidis, Premonition net, a multi-timeline transformer network architecture towards strawberry tabletop yield forecasting, Comput. Electron. Agric. 208 (2023) 107784.
[13] G. Pasqualino, A. Furnari, G. Signorello, G.M. Farinella, An unsupervised domain adaptation scheme for single-stage artwork recognition in cultural sites, Image Vis. Comput. 107 (2021) 104098.
[14] J.L. McClelland, B.L. McNaughton, R.C. O'Reilly, Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory, Psychol. Rev. 102 (3) (1995) 419.
[15] D. Kumaran, D. Hassabis, J.L. McClelland, What learning systems do intelligent agents need? Complementary learning systems theory updated, Trends Cogn. Sci. 20 (7) (2016) 512–534.
[16] T.L. Hayes, G.P. Krishnan, M. Bazhenov, H.T. Siegelmann, T.J. Sejnowski, C. Kanan, Replay in deep learning: Current approaches and missing biological elements, Neural Comput. 33 (11) (2021) 2908–2950.
[17] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7167–7176.
[18] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, J. Mach. Learn. Res. 13 (Mar) (2012) 723–773.
[19] M. Long, H. Zhu, J. Wang, M.I. Jordan, Deep transfer learning with joint adaptation networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 2208–2217.
[20] Y. Jiang, Y. Zhang, C. Lin, D. Wu, C.-T. Lin, EEG-based driver drowsiness estimation using an online multi-view and transfer TSK fuzzy system, IEEE Trans. Intell. Transp. Syst. 22 (3) (2020) 1752–1764.
[21] Y. Zhang, K. Xia, Y. Jiang, P. Qian, W. Cai, C. Qiu, L.K. Wee, D. Wu, Multimodality fusion & inductive knowledge transfer underlying non-sparse multi-kernel learning and distribution adaption, IEEE/ACM Trans. Comput. Biol. Bioinform. (2022).
[22] B. Sun, K. Saenko, Deep coral: Correlation alignment for deep domain adaptation, in: European Conference on Computer Vision, Springer, 2016, pp. 443–450.
[23] C. Chen, Z. Chen, B. Jiang, X. Jin, Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, 2019, pp. 3296–3303.
[24] D. Patel, K. Amin, A cross-domain semantic similarity measure and multi-source domain adaptation in sentiment analysis, in: 2022 International Conference on Augmented Intelligence and Sustainable Systems, ICAISS, IEEE, 2022, pp. 760–764.
[25] Z. Chen, C. Chen, X. Jin, Y. Liu, Z. Cheng, Deep joint two-stream Wasserstein auto-encoder and selective attention alignment for unsupervised domain adaptation, Neural Comput. Appl. (2019) 1–14.
[26] C.-Y. Lee, T. Batra, M.H. Baig, D. Ulbricht, Sliced wasserstein discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10285–10295.
[27] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep domain confusion: Maximizing for domain invariance, 2014, arXiv preprint arXiv:1412.3474.
[28] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, 2020, arXiv preprint arXiv:2002.05709.
[29] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big self-supervised models are strong semi-supervised learners, 2020, arXiv preprint arXiv:2006.10029.
[30] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[31] A. Durrant, G. Leontidis, Hyperspherically regularized networks for self-supervision, Image Vis. Comput. (2022) 104494.
[32] A. Bardes, J. Ponce, Y. LeCun, Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2021, arXiv preprint arXiv:2105.04906.
[33] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: International Conference on Machine Learning, PMLR, 2021, pp. 12310–12320.
[34] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent-a new approach to self-supervised learning, Adv. Neural Inf. Process. Syst. 33 (2020) 21271–21284.
[35] X. Chen, K. He, Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15750–15758.
[36] H. Ma, X. Li, X. Yuan, C. Zhao, Two-phase self-supervised pretraining for object re-identification, Knowl.-Based Syst. 261 (2023) 110220.
[37] A. Manová, A. Durrant, G. Leontidis, S-JEA: Stacked joint embedding architectures for self-supervised visual representation learning, 2023, arXiv preprint arXiv:2305.11701.
[38] M. Alkhalefi, G. Leontidis, M. Zhong, Semantic positive pairs for enhancing contrastive instance discrimination, 2023, arXiv preprint arXiv:2306.16122.

[39] A. Durrant, G. Leontidis, HMSN: Hyperbolic self-supervised learning by clustering with ideal prototypes, 2023, arXiv preprint arXiv:2305.10926.
[40] S. Tang, P. Su, D. Chen, W. Ouyang, Gradient regularized contrastive learning for continual domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 3, 2021, pp. 2665–2673.
[41] M. Schutera, F.M. Hafner, J. Abhau, V. Hagenmeyer, R. Mikut, M. Reischl, Cuepervision: self-supervised learning for continuous domain adaptation without catastrophic forgetting, Image Vis. Comput. 106 (2021) 104079, http://dx.doi.org/10.1016/j.imavis.2020.104079, URL https://www.sciencedirect.com/science/article/pii/S0262885620302110.
[42] G.I. Parisi, R. Kemker, J.L. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review, CoRR abs/1802.07569, 2018, URL http://arxiv.org/abs/1802.07569.
[43] Z. Li, D. Hoiem, Learning without forgetting, CoRR abs/1606.09282, 2016, URL http://arxiv.org/abs/1606.09282.
[44] J. Kirkpatrick, R. Pascanu, N.C. Rabinowitz, J. Veness, G. Desjardins, A.A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Overcoming catastrophic forgetting in neural networks,CoRR abs/1612.00796, 2016, URL http://arxiv.org/abs/1612.00796.
[45] H. Jung, J. Ju, M. Jung, J. Kim, Less-forgetting learning in deep neural networks, CoRR abs/1607.00122, 2016, URL http://arxiv.org/abs/1607.00122.
[46] S. Rebuffi, A. Kolesnikov, C.H. Lampert, iCaRL: Incremental classifier and representation learning, CoRR abs/1611.07725, 2016, URL http://arxiv.org/abs/1611.07725.
[47] J. Lee, J. Yoon, E. Yang, S.J. Hwang, Lifelong learning with dynamically expandable networks, CoRR abs/1708.01547, 2017, URL http://arxiv.org/abs/1708.01547.
[48] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324, http://dx.doi.org/10.1109/5.726791.
[49] R. Kemker, C. Kanan, FearNet: Brain-inspired model for incremental learning, CoRR abs/1711.10563, 2017, URL http://arxiv.org/abs/1711.10563.
[50] D. Lopez-Paz, M. Ranzato, Gradient episodic memory for continuum learning, CoRR abs/1706.08840, 2017, URL http://arxiv.org/abs/1706.08840,
[51] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, G. Tesauro, Learning to learn without forgetting by maximizing transfer and minimizing interference, CoRR abs/1810.11910, 2018, URL http://arxiv.org/abs/1810.11910.
[52] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P.K. Dokania, P.H. Torr, M. Ranzato, On tiny episodic memories in continual learning, 2019, arXiv preprint arXiv:1902.10486.
[53] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, C.H. Lampert, Icarl: Incremental classifier and representation learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2001–2010.
[54] Q. Pham, C. Liu, S.C.H. Hoi, DualNet: Continual learning, fast and slow, CoRR abs/2110.00175, 2021, URL https://arxiv.org/abs/2110.00175.
[55] R. Kemker, C. Kanan, Fearnet: Brain-inspired model for incremental learning, 2017, arXiv preprint arXiv:1711.10563.
[56] W. Chenshen, L. Herranz, L. Xialei, et al., Memory replay GANs: Learning to generate images from new categories without forgetting, in: The 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018, pp. 5966–5976.
[57] L. Pellegrini, G. Graffieti, V. Lomonaco, D. Maltoni, Latent replay for real-time continual learning, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2020, pp. 10203–10209.
[58] T.L. Hayes, K. Kafle, R. Shrestha, M. Acharya, C. Kanan, Remind your neural network to prevent catastrophic forgetting, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16, Springer, 2020, pp. 466–483.
[59] G.M. Van de Ven, H.T. Siegelmann, A.S. Tolias, Brain-inspired replay for continual learning with artificial neural networks, Nat. Commun. 11 (1) (2020) 4069.
[60] R. Volpi, D. Larlus, G. Rogez, Continual adaptation of visual representations via domain randomization and meta-learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4443–4453.
[61] M. Wulfmeier, A. Bewley, I. Posner, Incremental adversarial domain adaptation for continually changing environments, in: 2018 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2018, pp. 4489–4495.
[62] A. Bobu, E. Tzeng, J. Hoffman, T. Darrell, Adapting to continuously shifting domains, 2018.
[63] A. Bardes, J. Ponce, Y. LeCun, VICReg: Variance-invariance-covariance regularization for self-supervised learning, CoRR abs/2105.04906, 2021, URL https://arxiv.org/abs/2105.04906.
[64] J. O'Neill, B. Pleydell-Bouverie, D. Dupret, J. Csicsvari, Play it again: reactivation of waking experience and memory, Trends Neurosci. 33 (5) (2010) 220–229, http://dx.doi.org/10.1016/j.tins.2010.01.006, URL https://www.sciencedirect.com/science/article/pii/S0166223610000172.
[65] M.A. Wilson, B.L. McNaughton, Reactivation of hippocampal ensemble memories during sleep, Science 265 (5172) (1994) 676–679.
[66] D. Maltoni, V. Lomonaco, Continuous learning in single-incremental-task scenarios, CoRR abs/1806.08568, 2018, URL http://arxiv.org/abs/1806.08568.
[67] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, Y. Fu, Large scale incremental learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 374–382.
[68] H. Venkateswara, J. Eusebio, S. Chakraborty, S. Panchanathan, Deep hashing network for unsupervised domain adaptation, CoRR abs/1706.07522, 2017, URL http://arxiv.org/abs/1706.07522.
[69] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2066–2073.
[70] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[71] J.S. Denker, W. Gardner, H.P. Graf, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, H.S. Baird, I. Guyon, Neural network recognizer for hand-written zip code digits, in: Advances in Neural Information Processing Systems, Citeseer, 1989, pp. 323–331.
[72] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning, 2011.
[73] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains, in: ECCV, 2010.
[74] G. Griffin, A. Holub, P. Perona, Caltech-256 Object Category Dataset, California Institute of Technology, 2007.
[75] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[76] Y. You, I. Gitman, B. Ginsburg, Large batch training of convolutional networks, 2017, arXiv preprint arXiv:1708.03888.
[77] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv preprint arXiv:1807.03748.