
Article:
Ye, Fei and Bors, Adrian Gheorghe orcid.org/0000-0001-7838-0021 (2024) Lifelong Dual Generative Adversarial Nets Learning in Tandem. IEEE Transactions on Cybernetics. pp. 1353-1365. ISSN 2168-2267

https://doi.org/10.1109/TCYB.2023.3271388

Lifelong Dual Generative Adversarial Nets Learning in Tandem

Fei Ye and Adrian G. Bors, *IEEE Senior Member*
Department of Computer Science, University of York, York YO10 5GH, UK

***Abstract*—Continually capturing novel concepts without forgetting is one of the most critical functions sought for in artificial intelligence systems. However, even the most advanced deep learning networks are prone to quickly forgetting previously learnt knowledge after training with new data. The proposed Lifelong Dual Generative Adversarial Networks (LD-GANs) consists of two Generative Adversarial Networks (GANs), namely a Teacher and an Assistant teaching each other in tandem while successively learning a series of tasks. A single Discriminator is used to decide the realism of generated images by the dual GANs. A new training algorithm, called the Lifelong Self Knowledge Distillation (LSKD) is proposed for training the LD-GAN while learning each new task during lifelong learning (LLL). LSKD enables the transfer of knowledge from one more knowledgeable player to the other jointly with learning the information from a newly given dataset, within an adversarial playing game setting. In contrast to other lifelong learning models, LD-GANs is memory efficient and does not require freezing any parameters after learning each given task. Furthermore, we extend the LD-GANs to being the Teacher module in a Teacher-Student network for assimilating data representations across several domains during LLL. Experimental results indicate a better performance for the proposed framework in unsupervised lifelong representation learning when compared to other methods.**

***Index Terms*—Lifelong learning, Generative Adversarial Network (GAN), Representation Learning, Teacher-Student network.**

I. Introduction

Humans and living beings in general are able to learn during their entire life while artificial learning systems are far from achieving such capabilities [1]. Despite all recent achievements in the area of artificial intelligence, existing deep learning models are not enabled with lifelong learning abilities. Each time when a neural network is retrained on a new database, its old parameters are overwritten. This phenomenon is called catastrophic forgetting [2]. Lifelong learning capabilities would enable streaming the learning in artificial systems, which is essential in observing and analysing phenomena, semantic analysis of documents, surveillance, robot and unmanned vehicle control, or for adapting to changing environments among many other applications.

During lifelong learning (LLL), an agent or a model is trained on a series of tasks, where each task is associated with a different data domain. We assume that the model can only access the training samples from the current task while all previously learnt samples are unavailable during further training. In such a case, catastrophic forgetting represents a severe challenge for the model. The LLL aims to minimize the performance loss on the past learnt tasks while also achieving good performance on those learnt recently. Once all tasks have been learnt, we evaluate the generalization performance of a model on all testing sets, both from the past as well as those new. This paper mainly focuses on the lifelong generative modelling task, which is not well explored in other LLL research studies.

One of the solutions proposed to relieve catastrophic forgetting is to impose constraints on the parameters of the network [3], [4] using a regularization term, where the model's parameters important to past tasks will undergo smaller changes when learning new tasks. Another approach consists in increasing the number of neurons and network layers [5], or employing a specific learning metric [6]. Such approaches can preserve the optimal performance for all past tasks when updating model's parameters while adapting to new tasks as well. However, these methods are only used in the supervised learning setting, where the ground truth labels for each task are provided. Unlike the approaches mentioned above, in this paper, we consider unsupervised learning under the LLL setting, where class labels are not available, [5].

Generative Adversarial Nets (GANs) [7] represent one of the most popular methods in unsupervised learning, which provide good results in image synthesis [8], image-to-image translation [9], image dehazing [10], [11] and for learning interpretable representations [12], [13]. However, GANs only perform well on data samples originating from a single database. GANs can relieve catastrophic forgetting following self-supervised training, such as being retrained with generative replay samples. A new training set can be made up by combining data generated by a GAN and the newly available data [14]. Another approach consists in preserving some or old model's parameters [15] to be used later for training. LLL models based on GAN generators lack inference mechanisms, representing a critical problem in the lifelong unsupervised learning setting because they cannot capture complex data structures.

The proposed Lifelong Dual Generative Adversarial Networks (LD-GANs) addresses catastrophic forgetting by accumulating information through a dual Teacher-Assistant network, working in tandem, enabled by adversarial learning during the learning of a sequence of tasks. We introduce the Lifelong Self Knowledge Distillation (LSKD) which distils the information from a more knowledgeable generator to another one in a tandem process where a Teacher and an Assistant interchange their functions of teaching and learning. We also implement a Teacher-Student network, where the LD-GANs model represent the Teacher, used to train a lightweight

probabilistic generative model as a Student which acquire meaningful data representations.

The following contributions are brought in this paper:

- We propose the LD-GANs, a dual GAN model for learning successively a set of tasks, in tandem.
- A new lifelong training approach, namely the Lifelong Adversarial Knowledge Distillation (LSKD), represents an end-to-end memory efficient method for learning essential information from several tasks.
- We extend the proposed LD-GANs to a Teacher-Student network for enabling the online learning of statistical data representations while capturing both continuous and domain-specific generative factors across tasks.
- We introduce a new theoretical framework based on the Wasserstein distance, which provides new insights into the forgetting behaviour of the Student module when learning several tasks.

The rest of paper consists of Section II, which outlines the main approaches in the area of lifelong learning, Section III introduces the proposed Lifelong Dual Generative Adversarial Nets, while in Section IV we provide the theoretical analysis for the forgetting behaviour of the proposed approach. Finally, Section V contains the experimental results and their discussion, while the conclusions are drawn in Section VI.

II. Related works

Current research studies in lifelong learning can be grouped under three categories: memory-based systems [16], [17], [18], regularization-based [3], [4] and using architecture expansion [2], [5]. Memory-based approaches usually utilize a small memory buffer to store past samples and these are merged with novel samples for learning a new task aiming to relieve forgetting. Meanwhile, generators such as the Variational Autoencoder (VAE) [19] or Generative Adversarial Network (GAN) [7] can be used for reproducing previously learnt data. These generative replay samples are then mixed with data corresponding to a new task making up together a training set. GANs can be used as generative replay networks (GRM) [20], but they lack inference mechanisms, which prevent their applicability to many down-stream tasks, including image reconstruction and interpolation. Meanwhile, the Variational Autoencoder with Shared Embeddings (VASE) [21] aims to learn disentangled representations under lifelong learning. To enable learning meaningful latent variables across multiple domains, VASE introduces a new loss function based on the Minimum Description Length (MDL) principle [22], which progressively increases the representational capacity to accommodate learning new data. More recently, Ramapuram *et al.* [23] proposed the Lifelong Generative Modeling (LGM) which employs VAEs for two networks teaching each other as a Teacher and Student. Kuzina *et al.* [24] introduced the Boosting Approach for continual learning of VAE (BooVAE), which learns the approximation of the aggregated posterior as a prior for each given task. BooVAE uses the trainable pseudo-inputs as the parameters corresponding to the approximation of the posterior, thus preserving the past knowledge. Since VAEs have inference mechanisms, they can model cross-domain representations over several tasks. However, VAEs tend to produce relatively low-quality data, such as blurred images [25], [26], which negatively affects their ability to reproduce past information.

Regularization methods usually introduce an additional term in the optimization function, penalizing parameter changes when learning new tasks [27]. Kirkpatrick *et al.* [3] propose the Elastic Weight Consolidation (EWC), which employs a quadratic penalty on the difference between the parameters for the old and new tasks, aiming to minimize the change on the previously learnt parameters when learning a new task. However, one drawback of EWC is the growing computational complexity of learning a long sequence of tasks. This problem is addressed by Schwarz *et al.* [28], which only regulates the model updating to more recent tasks. The EWC was further improved in terms of reducing the computational cost by using a single diagonal Fisher matrix to preserve the information of all previously learned tasks, which is then updated using the moving average [29]. More recently, regularization approaches have been developed based on the Bayesian Inference framework. For instance, Nguyen *et al.* [27] introduced a new continual learning framework called the Variational Continual Learning (VCL), which employs the Bayesian principle to overcome forgetting. However, VCL requires to store past samples during inference. Hongjoon *et al.* [30] addressed the drawbacks of VCL, including the computation time and space complexity, by introducing two additional regularisation terms that preserve old knowledge by freezing important parameters and allocating the remaining capacity to tackle a new task. Chen *et al.* [31] used for text and images a hybrid approach for continual learning combining the advantages of the stochastic gradient Markov chain Monte Carlo and variance inference.

In Federated Learning [32], a collaboratively set of networks was used for training a model, while a GAN was employed as a Teacher in the Lifelong Teacher-Student [17]. In another group of approaches, Coupled GANs (CoGANs) [33] consist of a pair of GANs, where each generator and discriminator shares some of its weights with another generator and discriminator. DualGANs [34], has some structural similarities with CoGANs, while aiming to learn image-to-image correspondences, unlike the joint data distributions learnt in CoGANs. The Twin Auxiliary Classifiers GANs (TAC-GAN) [35] enforces data diversity by considering classifiers when interacting with a GAN's generator and discriminator.

The LD-GANs, introduced in this research study, has completely different characteristics from the other coupled approaches. Firstly, other coupled approaches update all their generators during training while LD-GANs trains only one generator for each task at a time. Secondly, LD-GANs transfers knowledge between the two generators using the proposed Lifelong Self Knowledge Distillation (LSKD) training algorithm. Thirdly, LD-GANs is able to learn new tasks without forgetting. Finally, the proposed LD-GANs can also provide higher-quality underlying data representation transfer from the Teacher to a Student model and consequently learn many informative latent representations over time.

III. Lifelong Dual Generative Adversarial Networks

In this section, we introduce the Lifelong Dual Generative Adversarial Networks (LD-GANs), which is afterwards extended into a Teacher-Student network.

A. Problem definition

Let us assume a set of N databases, each associated with a task, $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_N\}$. A lifelong model aims to be able to learn and use the probabilistic representations of all tasks $\mathcal{T}$ at any given time. Most existing lifelong learning approaches would normally employ a classifier $f(\mathbf{y}|\mathbf{x})$ which minimizes the empirical loss across all given tasks from $\mathcal{T}$. However, in practice it is impossible to acquire labels for a larger-scale database since it is time-consuming and requires extensive data annotation work. The problem of unsupervised learning under the lifelong setting has the advantage that the learner or agent does not have access to external supervision signals, including class labels or regression targets. Under this setting, the learning goal is to model a set of generative factors (latent variables) $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$, which are shared between different domains and can be used to explain their underlying data representations following knowledge distillation.

B. Learning a single task

The LD-GANs model contains two generators and a discriminator. One of the generators is called the Teacher while the other is the Assistant and their function is switched whenever LD-GANs is learning a new task. Let as consider a latent vector space $\mathcal{Z}$, defined by the random variable $\mathbf{z} \in \mathcal{Z}$, which is defined by a Normal distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The Teacher and Assistant, implemented by identical neural network structures $G_{\theta_T}(\mathbf{z})$ and $G_{\theta_A}(\mathbf{z})$, are used for generating the data $\mathbf{x}'_t$ and $\mathbf{x}'_a$, considered as images in the experiments, using $\mathbf{z}$ as input. When learning the first given task, the goal of LD-GANs is similar to that of a GAN [7] and in this study we consider minimizing the Wasserstein (Earth-Mover) distance as a probabilistic distance, [36], [37]:

$$\min_{G_{\theta_T}, G_{\theta_A}} \max_{D \in \Theta} \{\underbrace{\mathbb{E}_{\mathbf{x}^1 \sim p(\mathbf{x}^1)}[D(\mathbf{x}^1)] - \mathbb{E}_{\mathbf{x}'_t \sim p(\mathbf{x}_T)}[D(\mathbf{x}'_t)]}_{\text{Teacher optimization}} + \underbrace{\mathbb{E}_{\mathbf{x}^1 \sim p(\mathbf{x}^1)}[D(\mathbf{x}^1)] - \mathbb{E}_{\mathbf{x}'_a \sim p(\mathbf{x}_A)}[D(\mathbf{x}'_a)]}_{\text{Assistant optimization}}\}. \tag{1}$$

where $p(\mathbf{x}^1)$ as the probabilistic representation of the task defined by the first database, $p(\mathbf{x}_T)$ and $p(\mathbf{x}_A)$ represent the generator distributions for the Teacher $G_{\theta_T}(\mathbf{z})$ and Assistant $G_{\theta_A}(\mathbf{z})$, respectively, $D(\cdot)$ is the discriminant, and Θ define a set of 1-Lipschitz functions. We introduce a gradient penalty term (momentum) [38], enforcing the Lipschitz constraint:

$$\begin{aligned}\min_{G_{\theta_T}, G_{\theta_A}} \max_{D \in \Theta} \{&\mathbb{E}_{\mathbf{x}^1 \sim p(\mathbf{x}^1)}[D(\mathbf{x}^1)] - \mathbb{E}_{\mathbf{x}'_t \sim p(\mathbf{x}_T)}[D(\mathbf{x}'_t)] \\ &+ \lambda \mathbb{E}_{\tilde{\mathbf{x}}_t \sim \mathbb{P}_{\tilde{\mathbf{x}}_T}}[(\|\nabla_{\tilde{\mathbf{x}}_t} D(\tilde{\mathbf{x}}_t)\|_2 - 1)^2] \\ &+ \mathbb{E}_{\mathbf{x}^1 \sim p(\mathbf{x}^1)}[D(\mathbf{x}^1)] - \mathbb{E}_{\mathbf{x}'_a \sim p(\mathbf{x}_A)}[D(\mathbf{x}'_a)] \\ &+ \lambda \mathbb{E}_{\tilde{\mathbf{x}}_a \sim \mathbb{P}_{\tilde{\mathbf{x}}_A}}[(\|\nabla_{\tilde{\mathbf{x}}_a} D(\tilde{\mathbf{x}}_a)\|_2 - 1)^2]\},\end{aligned} \tag{2}$$

where $\mathbb{P}_{\tilde{\mathbf{x}}_T}$ and $\mathbb{P}_{\tilde{\mathbf{x}}_A}$ are joint distributions defining data interpolating samples from $p(\mathbf{x}^1)$ and pairs of real data, sampled from $p(\mathbf{x}_T)$ and $p(\mathbf{x}_A)$, generated by the Teacher and Assistant networks, respectively. The structure of the LD-GANs network and its training is shown in Fig. 1.

C. Lifelong Self Knowledge Distillation

Knowledge distillation learning assumes that a classifier is trained on the predictions of another classifier [39], [40]. For improving the performance, some recent studies propose employing an ensemble of networks [41], [42], which mixes the distributions of the predictions from an ensemble. However, these methods use real data and class information from a single domain, representing a severe challenge for the general application of the lifelong learning setting. In this research study, we introduce a new approach for knowledge transfer in LD-GANs, namely Lifelong Self Knowledge Distillation (LSKD). Through LSKD we employ one of the generators to be a Teacher and use its generated data for training another generator, called the Assistant. Then, following the learning of an initial dataset, we freeze the Teacher's parameters during the learning of a second database. Meanwhile, data from a second database is mixed with the data generated by the Teacher and together are used to train the Assistant. When learning each additional new task, the roles of the Teacher and Assistant are exchanged during learning in LSKD, with both GAN generators becoming alternatively the Teacher and the Assistant, respectively. The LSKD objective function is defined as:

$$\begin{aligned}\min_{G} \max_{D \in \Theta} \{&\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}^k) p(\mathbf{x}_T^{k-1})}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{x}' \sim p(\mathbf{x}_A^k)}[D(\mathbf{x}')] \\ &+ \lambda \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathbb{P}_{\tilde{\mathbf{x}}_A}}[(\|\nabla_{\tilde{\mathbf{x}}} D(\tilde{\mathbf{x}})\|_2 - 1)^2]\},\end{aligned} \tag{3}$$

where the data $\mathbf{x}$ are uniformly sampled from the probabilistic representation $p(\mathbf{x}^k)$ of the t-th task and the Teacher's distribution $p(\mathbf{x}_T^{k-1})$, after being trained on the $(k-1)$-th task. The sample $\mathbf{x}'$ is drawn from $p(\mathbf{x}_A^k)$ characterizing the data generated by the Assistant $G_{\theta_A^k}(\mathbf{z})$ following its training on the k-th task. The proposed LSKD training algorithm has multiple advantages over other lifelong learning methods [21], [43]. Firstly, LSKD is memory efficient because it does not require to know the data from all past databases [20]. Secondly, the model's parameters, or even a subset of these parameters, do not have to be preserved, as is the case for many other generative replay methods [21], [43]. The description of LSKD is provided in Algorithm 1.

D. Training a Student network for representation learning

In the following we extend LD-GANs to be used in a lifelong Teacher-Student framework [17]. The Teacher is implemented by the two GAN generators from LD-GANs, while the Student component is represented by a latent variable generative model $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$. The marginal likelihood of $p(\mathbf{x}, \mathbf{z})$ is intractable, requiring the integration over the entire latent variable space $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$. Instead,

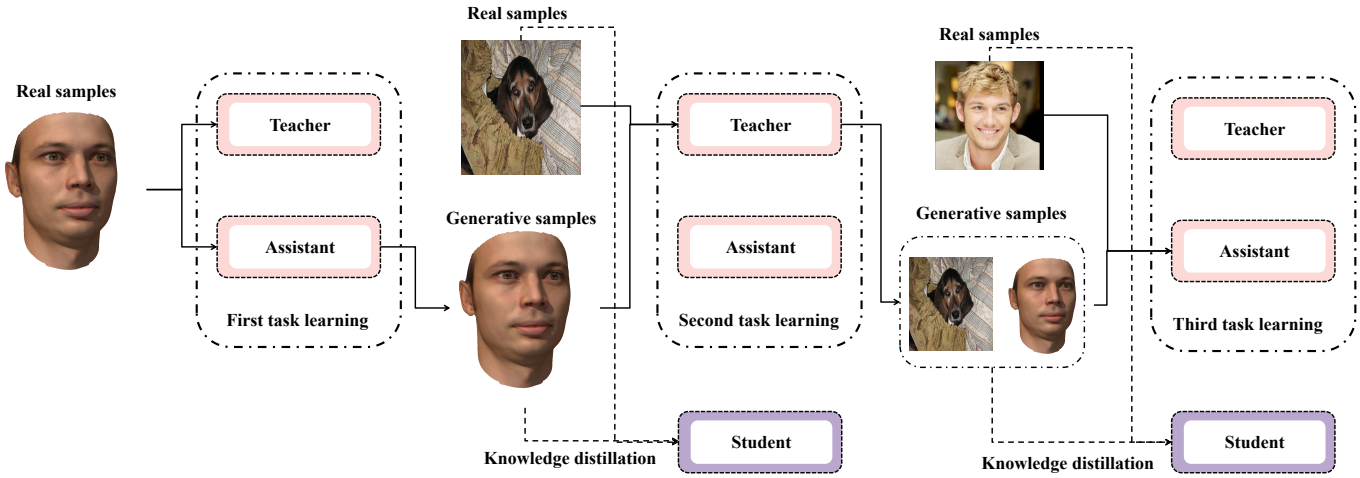


Fig. 1. The lifelong learning flow for LD-GANs. During the learning of the first task, both GAN generators, the Teacher and Assistant, are trained. Afterwards, the Teacher and Assistant teach each other alternatively, exchanging their roles after learning each task. Meanwhile, within the Teacher-Student architecture, the Student accumulates the generative data representations across domains over time.

we maximize the evidence lower bound (ELBO) on the sample log-likelihood, as in the VAE inference, [19]:

$$\begin{aligned}\log p(\mathbf{x}) \geq \mathbb{E}_{\mathbf{z}\sim q_\varepsilon(\mathbf{z}\,|\,\mathbf{x})}[\log p_\omega(\mathbf{x}\,|\,\mathbf{z})] \\ - D_{KL}[q_\varepsilon(\mathbf{z}|\mathbf{x})\,||\,p(\mathbf{z})] = \mathcal{L}_{\text{VAE}}(\omega,\varepsilon),\end{aligned} \tag{4}$$

where $p_\omega(\mathbf{x}\,|\,\mathbf{z})$ is the decoder and $q_\varepsilon(\mathbf{z}\,|\,\mathbf{x})$ is an inference network of prior parameters $\{\boldsymbol{\mu},\boldsymbol{\sigma}\}$, characterizing the mean and variance of a Gaussian implemented by the last network' layer, while D_{KL} represents the Kullback-Leibler (KL) divergence. The latent vector $\mathbf{z}$ is sampled by means of the reparametrisation trick $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\gamma} \odot \boldsymbol{\sigma}$, where $\boldsymbol{\gamma}$ is a random vector drawn from $\mathcal{N}(\mathbf{0},\mathbf{I})$ and $\odot$ is the element-wise product.

In order to learn cross-domain representations under the lifelong learning framework, we transfer the knowledge from the most knowledgeable generator, out of the two from the LD-GANs, to the Student network. Therefore, we define the objective function according to the following Lemma:

Lemma 1: Let us consider that $\mathbf{x}^k$ and $\mathbf{x}_Q^{k-1}$ represent samples from the data distribution characterizing the k-th task and also which are synthesized by the generator distribution (Teacher), respectively. We define a new distillation loss function guaranteeing a lower bound to the data log-likelihood over the joint variables $\{\mathbf{x}^k, \mathbf{x}_Q^{k-1}\}$ at the k-th task learning:

$$\begin{aligned}&\log[p(\mathbf{x}^k)p(\mathbf{x}_Q^{k-1})] \geq \mathcal{L}_{stu}(\omega,\varepsilon) = \\ &\underbrace{\mathbb{E}_{\mathbf{z}\sim q_\varepsilon(\mathbf{z}|\mathbf{x}^k)}[\log p_\omega(\mathbf{x}\,|\,\mathbf{z})] - D_{KL}[q_\varepsilon(\mathbf{z}\,|\,\mathbf{x}^k)\,||\,p(\mathbf{z})]}_{\text{Loss on samples drawn from the } k\text{-th task}} \\ &+ \underbrace{\mathbb{E}_{\mathbf{z}\sim q_\varepsilon(\mathbf{z}\,|\,\mathbf{x}')}[\log p_\omega(\mathbf{x}\,|\,\mathbf{z})] - D_{KL}[q_\varepsilon(\mathbf{z}|\mathbf{x}')\,||\,p(\mathbf{z})]}_{\text{KD loss}}.\end{aligned} \tag{5}$$

where $p(\mathbf{x}_Q^{k-1})$ can be either $p(\mathbf{x}_A^{k-1})$ or $p(\mathbf{x}_T^{k-1})$, depending on which one is more knowledgeable when learning the k-th task, and $\mathbf{x}'$ represents its generated data samples.

Proof. We assume that $(\mathbf{x}^k, \mathbf{x}_Q^{k-1}) \sim p(\mathbf{x}^k, \mathbf{x}_Q^{k-1})$ is a pair of samples drawn from the joint distribution. We also know that $p(\mathbf{x}^k)$ is independent from $p(\mathbf{x}_Q^{k-1})$. Then we define a latent variable model $p_\omega(\mathbf{x}^k, \mathbf{x}_Q^{k-1}, \mathbf{z}^k, \mathbf{z}_Q^{k-1}) = p_\omega(\mathbf{x}^k, \mathbf{x}_Q^{k-1}\,|\,\mathbf{z}^k, \mathbf{z}_Q^{k-1})p(\mathbf{z}^k, \mathbf{z}_Q^{k-1})$, with the marginal log distribution defined as:

$$\log p_\omega(\mathbf{x}^k, \mathbf{x}_Q^{k-1}) = \log \iint p(\mathbf{x}^k, \mathbf{x}_Q^{k-1}, \mathbf{z}^k, \mathbf{z}_Q^{k-1}) d\mathbf{z}^k d\mathbf{z}_Q^{k-1}. \tag{6}$$

We further assume that $\mathbf{x}_Q^{k-1}$ is also independent from $\mathbf{x}^k$ and $\mathbf{z}^k$ from $\mathbf{z}_Q^{k-1}$. We decompose Eq. (6) as:

$$\begin{aligned}\log p_\omega(\mathbf{x}^k, \mathbf{x}_Q^{k-1}) = \log \int p(\mathbf{x}^k, \mathbf{z}^k) d\mathbf{z}^k \\ \cdot p(\mathbf{x}_Q^{k-1}, \mathbf{z}_Q^{k-1}) d\mathbf{z}_Q^{k-1}.\end{aligned} \tag{7}$$

Then we can further consider :

$$\begin{aligned}\log p_\omega(\mathbf{x}^k, \mathbf{x}_Q^{k-1}) &= \log \int p(\mathbf{x}^k, \mathbf{z}^k) \frac{q_\varepsilon(\mathbf{z}|\mathbf{x}^k)}{q_\varepsilon(\mathbf{z}|\mathbf{x}^k)} d\mathbf{z}^k \\ &\cdot p(\mathbf{x}_Q^{k-1}, \mathbf{z}_Q^{k-1}) \frac{q_\varepsilon(\mathbf{z}_Q^{k+1}|\mathbf{x}_Q^{k+1})}{q_\varepsilon(\mathbf{z}_Q^{k+1}|\mathbf{x}_Q^{k+1})} d\mathbf{z}_Q^{k-1} \\ &= \log \left(\mathbb{E}_{q_\varepsilon(\mathbf{z}|\mathbf{x}^k)} \frac{p(\mathbf{x}^k, \mathbf{z}^k)}{q_\varepsilon(\mathbf{z}|\mathbf{x}^k)} \right. \\ &\left. \cdot \mathbb{E}_{q_\varepsilon(\mathbf{z}_Q^{k+1}|\mathbf{x}_Q^{k+1})} \frac{p(\mathbf{x}_Q^{k-1}, \mathbf{z}_Q^{k-1})}{q_\varepsilon(\mathbf{z}_Q^{k+1}|\mathbf{x}_Q^{k+1})} \right).\end{aligned} \tag{8}$$

By using the Jensens inequality, we have:

$$\begin{aligned}\log p_\omega(\mathbf{x}^k, \mathbf{x}_Q^{k-1}) &\geq \mathbb{E}_{\mathbf{z}^k\sim q_\varepsilon(\mathbf{z}^k\,|\,\mathbf{x}^k)} \log \left[\frac{p_\omega(\mathbf{x}^k, \mathbf{z}^k)}{q_\varepsilon(\mathbf{z}|\mathbf{x}^t)} \right] \\ &+ \mathbb{E}_{\mathbf{z}_Q^{k-1}\sim q_\varepsilon(\mathbf{z}_Q^{k-1}\,|\,\mathbf{x}_Q^{k-1})} \log \left[\frac{p_\omega(\mathbf{x}_Q^{k-1}, \mathbf{z}_Q^{k-1})}{q_\varepsilon(\mathbf{z}^{k+1_Q}|\mathbf{x}^{k+1_Q})} \right].\end{aligned} \tag{9}$$

Eventually, the above equation is rewritten as:

$$\begin{aligned}&\log p(\mathbf{x}^k)p(\mathbf{x}_Q^{k-1}) \geq \mathcal{L}_{stu}(\omega,\varepsilon) = \\ &\underbrace{\mathbb{E}_{\mathbf{z}\sim q_\varepsilon(\mathbf{z}\,|\,\mathbf{x}^k)}[\log p_\omega(\mathbf{x}\,|\,\mathbf{z})] - D_{KL}[q_\varepsilon(\mathbf{z}\,|\,\mathbf{x}^k)\,||\,p(\mathbf{z})]}_{\text{Loss on samples drawn from the } k\text{-th task}} \\ &+ \underbrace{\mathbb{E}_{\mathbf{z}\sim q_\varepsilon(\mathbf{z}\,|\,\mathbf{x}')}[\log p_\omega(\mathbf{x}\,|\,\mathbf{z})] - D_{KL}[q_\varepsilon(\mathbf{z}\,|\,\mathbf{x}')\,||\,p(\mathbf{z})]}_{\text{KD loss}},\end{aligned} \tag{10}$$

where we simply use the same $\mathbf{z}$ for all latent variables since we use a single inference model and $\mathbf{x}'$ replaces $\mathbf{x}_Q^{k-1}$ for simplicity. More importantly, this choice can allow the inference model to automatically capture both shared and domain-specific generative factors in the same latent space.

The loss defined by Eq. (10), corresponding to maximizing ELBO on the joint sample log-likelihood, is used to train the Student network when learning the k-th task, $k > 1$. In practice, the Student network learning is synchronized with training the LD-GANs in each task and is the only part of the model trained when learning the last task, providing a flexible training manner for the proposed Teacher-Student framework.

Algorithm: We provide the pseudocode in Algorithm 1, which is used for training the proposed Teacher-Student framework and can be summarized into three steps:

Step 1. The first task learning: We draw training samples from the first task, which are used for training both the Teacher and Assistant using Eq. (1). Using the first dataset, we also train the Student using Eq. (4).

Step 2. The subsequent task learning: If the Teacher learns more tasks than the Assistant at the i-th task learning, we fix the Teacher and treat it as the generative replay network. Then we incorporate generative replay samples from the Teacher and real samples from the i-th task, which are used for training the Assistant using Eq. (3). If the Assistant learns more tasks than the Teacher, we interchange their roles aiming for the model to accumulate more knowledge during the lifelong learning.

Step 3. The knowledge distillation: We transfer the information from a more knowledgeable Teacher component (Teacher or Assistant) to the Student while allowing the Student to learn novel samples from the new task as well, using Eq. (10).

Algorithm 1: The unsupervised learning for LD-GANs

Input: A sequence of datasets;
Input: The number of tasks (n);
Input: The number of iterations (m);
Output: The parameters of the model;

for $i < n$ **do**
  for $j < m$ **do**
    Teacher learning;
    if *i==0* **then**
      $\mathbf{x} \sim p(\mathbf{x}^i)$;
      Train $G_{\theta_T}(\mathbf{z})$ and $G_{\theta_A}(\mathbf{z})$ on $\mathbf{x}$ by Eq. (2);
      Train the Student on $\mathbf{x}$ using Eq. (4);
    end
    else
      if *i mod 2 == 0* **then**
        $\mathbf{x} \sim p(\mathbf{x}^i)p(\mathbf{x}_A^k)$;
        Train $G_{\theta_T}(\mathbf{z})$ while fixing $G_{\theta_A}(\mathbf{z})$ on $\mathbf{x}$;
      end
      else
        $\mathbf{x} \sim p(\mathbf{x}^i)p(\mathbf{x}_T^k)$;
        Train $G_{\theta_A}(\mathbf{z})$ while fixing $G_{\theta_T}(\mathbf{z})$ on $\mathbf{x}$;
      end
    end
    Knowledge distillation;
    if *i mod 2 == 0* **then**
      $\mathbf{x} \sim p(\mathbf{x}^i)p(\mathbf{x}_A^k)$;
    end
    else
      $\mathbf{x} \sim p(\mathbf{x}^i)p(\mathbf{x}_T^k)$;
    end
    Train the Student using Eq. (10) on $\mathbf{x}$;
  end
end

E. Learning disentangled representations over time

In the following, we enable the Student network to learn disentangled representations across domains. The Total Correlation (TC) term has been used in various VAE frameworks [44], [45] to encourage learning disentangled representations for a single database. However, these approaches require an extra sampling process [45], or an additional discriminator network which is used to estimate the TC term [44]. In this research study, we simply penalize the Kullback-Leibler (KL) divergence between the posterior and prior distributions [22] in order to encourage disentanglement in the latent variables. By enforcing disentanglement, we model data properties over continuously learning several tasks. The resulting disentangled loss for the Student model is defined as:

$$\begin{aligned}\log[p(\mathbf{x}^k)p(\mathbf{x}_Q^{k-1})] \geq\ & \mathbb{E}_{\mathbf{z}\sim q_\varepsilon(\mathbf{z}\,|\,\mathbf{x}^k)}[\log p_\omega(\mathbf{x}\,|\,\mathbf{z})] \\ & - \beta D_{KL}[q_\varepsilon(\mathbf{z}\,|\,\mathbf{x}^k)\,||\,p(\mathbf{z})] \\ & + \mathbb{E}_{\mathbf{z}\sim q_\varepsilon(\mathbf{z}\,|\,\mathbf{x}')}[\log p_\omega(\mathbf{x}\,|\,\mathbf{z})] \\ & - \beta D_{KL}[q_\varepsilon(\mathbf{z}\,|\,\mathbf{x}')\,||\,p(\mathbf{z})] = \mathcal{L}_{Dis}(\omega,\varepsilon).\end{aligned} \tag{11}$$

For $\beta = 1$, Eq. (11) corresponds to Eq. (10) and a large β leads to more independent latent variables while decreasing the reconstruction quality, [46].

F. Supervised Learning

Although the proposed LD-GANs mainly focuses on unsupervised learning, we show that LD-GANs can be extended to supervised learning with minimal modifications. To implement this goal, we introduce a classifier for the Teacher and Assistant, represented by f_{φ_t} and f_{φ_a}, respectively. During the learning of the first task, we train jointly f_{φ_t} and f_{φ_a} on the first dataset by using the loss functions:

$$\mathcal{L}_{t1} = \frac{1}{n}\sum_{i=1}^{n}\{\mathcal{L}_{ce}(f_{\varphi_t}(\mathbf{x}_i^1), \mathbf{y}_i^1))\}, \tag{12}$$

$$\mathcal{L}_{a1} = \frac{1}{n}\sum_{i=1}^{n}\{\mathcal{L}_{ce}(f_{\varphi_a}(\mathbf{x}_i^1), \mathbf{y}_i^1))\}, \tag{13}$$

where $\mathcal{L}_{ce}(\cdot)$ is the cross-entropy loss and $\{\mathbf{x}_i^1, \mathbf{y}_i^1\}$ is the i-th paired samples, each representing a sample and its corresponding class, $i = 1, \ldots, n$, drawn from the distribution $\{p(\mathbf{x}^1), p(\mathbf{y}^1)\}$ from the first dataset. When learning a new task, such as the j-th task, we train the classifier alternately on a joint dataset which consists of real samples from the j-th task and generated samples drawn from one of the two GANs, whichever is the Teacher at that learning stage. After the lifelong learning process is finished, we choose the classifier which has learnt more tasks for evaluation.

(a) LD-GAN generations.

(b) Real testing samples.

(c) LD-GAN reconstructions.

(d) LTS Generations.

(e) LTS reconstructions.

Fig. 2. Generations and reconstructions after CCCSSM lifelong learning.

IV. Theoretical analysis for the forgetting behaviour

In this section, we study the forgetting behaviour of LD-GANs by deriving its learning bounds based on the Wasserstein distance [47].

A. Preliminary

Definition 1: (Data distributions.) Let Q_i and P_i represent the distribution for the testing and training sets of the i-th task, respectively.

Definition 2: (Approximate distribution.) Let $\mathbb{P}_j$ represent the sample distribution drawn from the data generated by the Teacher after learning the j-th task. We assume that we have an optimal task labelling function F_l which receives a sample and returns its true task label. We can form the approximate distribution for a certain task (i-th task) at the j-th task learning by the sampling process :

$$\mathbf{X}' = \{\mathbf{x}'_j \sim \mathbb{P}_j \,|\, \mathrm{F}_l(\mathbf{x}'_j) = i, j = 1, \ldots, n\}, \tag{14}$$

where n is the total number of samples. Let $\mathbb{P}^i_{(j,j-i)}$ denote the probability distribution of samples from Eq. (14), where $j-i$ represent the number of GRM processes needed for generating the data for i-th task when training with the j-th task. We also use $\mathbb{P}^i_{(j,0)}$ and $\mathbb{P}^i_{(j,-1)}$ to represent P_i and Q_i, respectively.

Definition 3: (Risks.) Let $\mathcal{H}$ be a class of hypotheses, and $h \in \mathcal{H}$ a hypothesis implemented by the Student module. For a given domain P_i, we define the risk as:

$$\mathcal{R}(h, P_i) = \frac{1}{m} \sum_{k=1}^{m} \mathcal{L}(h(\mathbf{x}_k), \mathbf{x}_k), \tag{15}$$

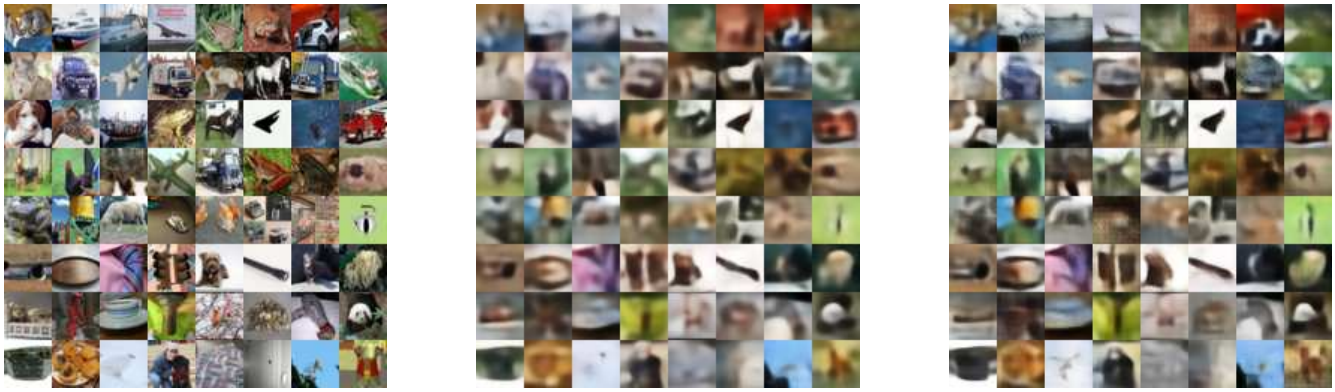
where $\mathcal{L}$ is the loss function implemented by the reconstruction error and $h(\cdot)$ returns the reconstruction. Each $\mathbf{x}_k$ is drawn from P_i and m is the total number of samples.

B. Forgetting analysis

We first derive the risk bound between the two fixed domains Q_i and P_i based on these definitions.

Theorem 1: Let us consider U_{x1} and U_{x2} be two sample populations of sizes N_S and N_T, drawn from two domains Q_i and P_i, respectively, while $\tilde{Q}_i$ and $\tilde{P}_i$ are the empirical probabilistic representations for U_{x1} and U_{x2}, respectively. With the probability of $1-u$, we have the following generalization bound (GB):

$$\begin{aligned}\mathcal{R}\left(h, Q_i\right) \leq \mathcal{R}\left(h, P_i\right) + \mathrm{W}_1\left(\tilde{Q}_i, \tilde{P}_i\right) \\ + \sqrt{2\log\left(\frac{1}{u}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right) + \mathcal{R}_\lambda,\end{aligned} \tag{16}$$



(a) Real samples. (b) CURL [48] reconstructions. (c) LD-GANs reconstructions.

(d) CURL generations. (e) LD-GANs generations.

Fig. 3. Image generations and reconstructions after the lifelong learning of CCCSSM sequence of tasks.

where W_1 is the Wasserstein distance, $\sqrt{2} > \varsigma' > 0$ and we have the combined error

$$\mathcal{R}_\lambda = \mathcal{R}(h^*, P_i) + \mathcal{R}(h^*, Q_i), \tag{17}$$

achieved by the optimal hypothesis h^* that minimizes this error. The detailed proof is provided in [47].

Theorem 1 defines the risk bound that measures the generalization performance on Q_i, achieved by the Student model h trained on P_i. In the following, we extend the results from Theorem 1 for deriving a risk bound that measures the generalization of the model when learning several tasks.

Theorem 2: Let us consider the proposed LD-GANs model when learning the j-th task. We derive the risk bound for a certain i-th task, learnt by the model in the past, as:

$$\begin{aligned}\mathcal{R}\left(h, Q_i\right) &\leq \mathcal{R}\left(h, \mathbb{P}^i_{j,j-i}\right) + \mathrm{W}_1\left(\tilde{Q}_i, \tilde{\mathbb{P}}^i_{j,j-i}\right) \\ &+ \sqrt{2\log\left(\frac{1}{u}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right) + \mathcal{R}_\lambda(Q_i, \mathbb{P}^i_{j,j-i}),\end{aligned} \tag{18}$$

where $\mathbb{P}^i_{j,j-1}$ is the data probabilistic representation from the j-th task, achieved through the generation process of the model and $\tilde{\mathbb{P}}^i_{j,j-i}$ its empirical probabilistic representations and

$$\mathcal{R}_\lambda(Q_i, \mathbb{P}^i_{j,j-1}) = \mathcal{R}(h^*, Q_i) + \mathcal{R}(h^*, \mathbb{P}^i_{j,j-i}). \tag{19}$$

Remark. From Theorem 2, we have several observations.

- The generalization performance of h is relying on the Wasserstein distance between the empirical distribution $\tilde{Q}_i$ and the approximate distribution $\tilde{\mathbb{P}}^i_{j,j-i}$.
- By learning more tasks (j is increased), the term $\mathrm{W}_1(\tilde{Q}_i, \tilde{\mathbb{P}}^i_{j,j-i})$ increases, leading to the degenerated performance on the target domain Q_i.

In the following, we extend Theorem 2 to derive a lemma demonstrating how accumulated errors contribute to forgetting.

Lemma 2: Let us consider that LD-GANs model is trained with the j-th task and we evaluate its performance for a certain learnt task, such as the i-th task, $i < j$, considering that the learning process has undergone repeatedly Generative Reply Mechanisms (GRMs), each time when training on a new database. The risk bound is derived as:

$$\begin{aligned}\mathcal{R}\left(h, Q_i\right) &\leq \mathcal{R}\left(h, \mathbb{P}^i_{j,j-i}\right) + \sum_{k=0}^{j-i-1}\left\{\mathrm{W}_1\left(\tilde{\mathbb{P}}^i_{j,k}, \tilde{\mathbb{P}}^i_{j,k+1}\right)\right. \\ &\left. + \mathcal{R}_\lambda(\tilde{\mathbb{P}}^i_{j,k}, \tilde{\mathbb{P}}^i_{j,k+1})\right\} \\ &+ (k-j)\sqrt{2\log\left(\frac{1}{u}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right),\end{aligned} \tag{20}$$

Proof. Firstly, let $\mathbb{P}^i_{j,0}$ and $\mathbb{P}^i_{j,1}$ be the target and source domain, respectively. We can derive the risk bound according

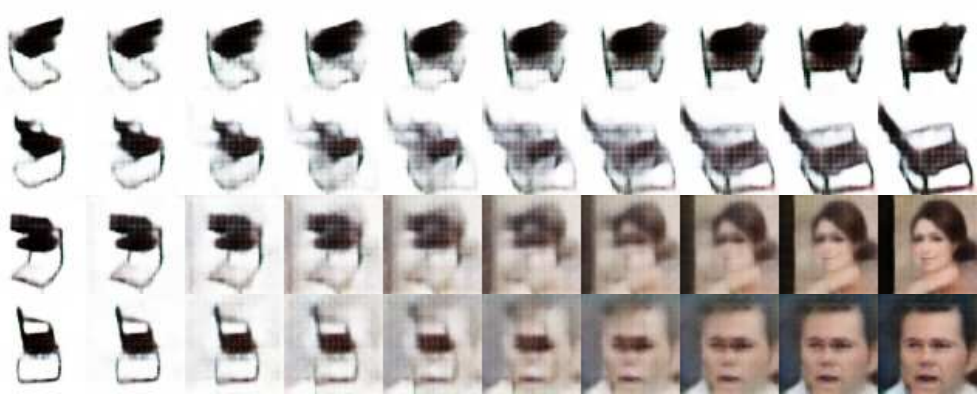


Fig. 4. Generative results when interpolating in the latent space of the Student network, under the lifelong LD-GANs Teacher-Student learning setting. The top two rows show interpolations between images from the 3D-Chair dataset, while the bottom two show interpolated images between an image from 3D-Chair and another from CelebA databse.

to Theorem 2:

$$\mathcal{R}\left(h, \mathbb{P}^i_{j,0}\right) \leq \mathcal{R}\left(h, \mathbb{P}^i_{j,1}\right) + \mathrm{W}_1\left(\tilde{\mathbb{P}}^i_{j,0}, \tilde{\mathbb{P}}^i_{j,1}\right) + \sqrt{2\log\left(\frac{1}{u}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right) + \mathcal{R}_\lambda(\tilde{\mathbb{P}}^i_{j,0}, \tilde{\mathbb{P}}^i_{j,1}). \tag{21}$$

We then take $\mathbb{P}^i_{j,1}$ and $\mathbb{P}^i_{j,2}$ as the target and source domain and we have:

$$\mathcal{R}\left(h, \mathbb{P}^i_{j,1}\right) \leq \mathcal{R}\left(h, \mathbb{P}^i_{j,2}\right) + \mathrm{W}_1\left(\tilde{\mathbb{P}}^i_{j,1}, \tilde{\mathbb{P}}^i_{j,2}\right) + \sqrt{2\log\left(\frac{1}{u}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right) + \mathcal{R}_\lambda(\tilde{\mathbb{P}}^i_{j,1}, \tilde{\mathbb{P}}^i_{j,2}), \tag{22}$$

Following mathematical induction, we have the following bounds:

$$\begin{aligned}
&\mathcal{R}\left(h, \mathbb{P}^i_{j,2}\right) \leq \mathcal{R}\left(h, \mathbb{P}^i_{j,3}\right) + \mathrm{W}_1\left(\tilde{\mathbb{P}}^i_{j,2}, \tilde{\mathbb{P}}^i_{j,3}\right) \\
&+ \sqrt{2\log\left(\frac{1}{u}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right) + \mathcal{R}_\lambda(\tilde{\mathbb{P}}^i_{j,2}, \tilde{\mathbb{P}}^i_{j,3}) \\
&\cdots \\
&\mathcal{R}\left(h, \mathbb{P}^i_{j,j-i-1}\right) \leq \mathcal{R}\left(h, \mathbb{P}^i_{j,j-1}\right) + \mathrm{W}_1\left(\tilde{\mathbb{P}}^i_{j,j-i-1}, \tilde{\mathbb{P}}^i_{j,j-i}\right) \\
&+ \sqrt{2\log\left(\frac{1}{u}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right) \\
&+ \mathcal{R}_\lambda(\tilde{\mathbb{P}}^i_{j,j-1-i}, \tilde{\mathbb{P}}^i_{j,j-1}).
\end{aligned} \tag{23}$$

Then we sum up all the above expressions, resulting in:

$$\begin{aligned}
\mathcal{R}\left(h, Q_i\right) \leq \mathcal{R}\left(h, \mathbb{P}^i_{j,j-i}\right) + \sum_{k=0}^{j-i-1}\Big\{&\mathrm{W}_1\left(\tilde{\mathbb{P}}^i_{j,k}, \tilde{\mathbb{P}}^i_{j,k+1}\right) \\
&+ \mathcal{R}_\lambda(\tilde{\mathbb{P}}^i_{j,k}, \tilde{\mathbb{P}}^i_{j,k+1})\Big\} \\
&+ (k-j)\sqrt{2\log\left(\frac{1}{u}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right).
\end{aligned} \tag{24}$$

□

Remark. From Lemma 2, we have several observations.

- The second term from the right-hand side of Eq. (20) is increasing by accumulating errors during each task learning. When learning a growing number of tasks, the error of the model tends to increase.
- The model's performance on the early tasks (i is very small) has more accumulated errors, resulting in a degenerated performance.

In the following, we extend Lemma 2 to derive the risk bound for all tasks.

Theorem 3: Let us consider that LD-GANs is learning j-th task. We derive the risk bound for all previously learnt tasks as:

$$\begin{aligned}
\sum_{t=1}^{j}\Big\{\mathcal{R}\left(h, Q_t\right)\Big\} \leq \sum_{t=1}^{j}\Big\{&\mathcal{R}\left(h, \mathbb{P}^t_{j,j-i}\right) \\
&+ \sum_{k=0}^{j-t-1}\Big\{\mathrm{W}_1\left(\tilde{\mathbb{P}}^t_{j,k}, \tilde{\mathbb{P}}^t_{j,k+1}\right) + \mathcal{R}_\lambda(\tilde{\mathbb{P}}^t_{j,k}, \tilde{\mathbb{P}}^t_{j,k+1})\Big\} \\
&+ (k-j)\sqrt{2\log\left(\frac{1}{u}\right)/\varsigma'}\left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right)\Big\},
\end{aligned} \tag{25}$$

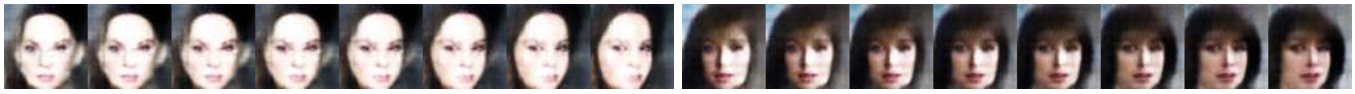
The proof results from summing up the risk bounds for all tasks, according to Lemma 2.

Remark. From Theorem 3, we have several observations.

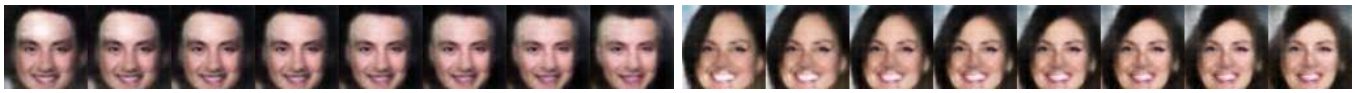
- The optimal performance can be achieved by minimizing the Wasserstein distance between the target and source distribution when learning every given task.
- In practice, when learning several different probabilistic representations, characterizing a variety of tasks, increases the Wasserstein distance when learning each new task, and the GAN model would eventually face mode collapse.
- The proposed LD-GANs model can relieve this challenge through balancing replay mechanisms through the LSKD training algorithm, as explained in Section III-C.

V. Experiments

In this section we provide the experimental results when evaluating the abilities of the proposed LD-GANs model, which consists of two GAN networks, representing the Teacher and Assistant, which alternatively teach each other after each task switch, for learning a succession of databases. One of



(a) Modelling the turning of a face.



(b) Changing the width of a face.

(c) Chair backside change.

(d) Chair size change.

Fig. 5. Disentanglement results following the LD-GANs learning of CelebA to 3D-Chair.

TABLE I
THE PERFORMANCE OF VARIOUS MODELS UNDER THE MSFIR LIFELONG LEARNING SETTING, WHERE THE RESULT FOR LIMIX-STU IS REPORTED FROM [2]

	SSIM					PSNR				
Datasets	LGM [23]	LD-GAN	BE-Stu [49]	LTS [17]	LIMix-Stu [2]	LGM [23]	LD-GAN	BE-Stu [49]	LTS [17]	LIMix-Stu [2]
MNIST	0.81	0.73	**0.86**	0.71	0.42	18.34	17.10	**20.16**	16.56	13.72
Fashion	0.40	0.54	0.41	**0.71**	0.37	10.63	11.91	12.33	**17.76**	8.81
SVHN	0.25	**0.72**	0.45	0.45	0.47	7.56	**17.77**	11.01	11.03	13.58
IFashion	0.30	**0.76**	0.60	0.75	0.43	7.27	**18.34**	15.46	18.05	14.17
RMNIST	**0.91**	0.90	0.90	0.89	0.43	**22.01**	21.64	21.58	21.31	14.18
Average	0.53	**0.73**	0.64	0.70	0.42	13.16	**17.35**	16.11	16.94	12.89

aims of LD-GANs is to learn meaningful and interpretable latent representations across domains. We also evaluate the performance of the proposed LD-GANs on several down stream tasks, including prediction and classification tasks. The source code is provided at https://github.com/dtuzi123/Lifelong-Dual-GAN.

A. Datasets and evaluation criteria

Datasets. We follow the learning setting from [2] which considers a sequence of five tasks including MNIST [50], SVHN [51], Fashion [52], InverseFashion (IFashion) and Rotated MNIST (RMNIST) databases. We name this learning setting as MSFIR. We also consider a sequence named CCCSSM of datasets, which contain images of higher complexity, including CelebA [53], CACD [54], CIFAR10 [55], Sub-ImageNet [56], SVHN and MNIST.

Baselines. Since this paper mainly focuses on the generative replay methods, we compare our approach with this category of lifelong learning approaches, including LGM [23] and MemoryGANs [57]. We also compare LD-GANs with Teacher-Student models including LTS [17] and BE-Stu [49]. We implement BE-Stu by using the BatchEnsemble [49] as the Teacher where each component is a VAE and the model shares parameters between components. The Student in BE-Stu is implemented by a VAE which is trained on the generated images by the Teacher. We also compare with LIMix-Stu [2] which uses the Teacher-Student framework.

Evaluation criteria. We adapt the criteria according to the unsupervised learning setting from [2], which includes the structural similarity index measure (SSIM) [58] and the Peak-Signal-to-Noise Ratio (PSNR) [58] as performance criteria.

B. The evaluation of the generative task

We evaluate the performance of the trained Student model on the image generation task. Firstly, we train our model under MSFIR, where the number of training epochs for learning each task is 20. We use the Stochastic Gradient Descent (SGD) algorithm with a learning rate of 0.0002 for training LD-GANs. After the training with all given tasks is finished, we evaluate the performance of various models on each testing dataset and report the results in Table I. From this table we observe that LD-GANs achieves the best results for both SSIM and PSNR criteria when compared with other methods. From Table I we can also observe that GAN-based models such as

TABLE II
THE PERFORMANCE OF VARIOUS MODELS UNDER THE MSFIR LIFELONG LEARNING SETTING.

	FID					IS				
Datasets	LGM [23]	LD-GAN	BE-Stu [49]	LTS [17]	LIMix-Stu [2]	LGM [23]	LD-GAN	BE-Stu [49]	LTS [17]	LIMix-Stu [2]
MNIST	82.62	81.15	36.28	75.80	104.94	3.12	3.08	3.97	2.29	1.93
Fashion	155.12	72.41	203.81	126.16	217.87	2.67	3.64	3.30	3.68	2.93
SVHN	187.07	167.06	322.87	180.15	286.89	2.06	2.14	2.37	2.37	2.84
IFashion ...	77.25	49.46	391.90	120.55	314.49	3.10	3.84	1.03	3.49	1.63
RMNIST	76.29	24.27	26.26	41.01	51.36	2.25	2.08	2.02	2.00	1.99
Average	115.67	**78.87**	196.22	108.73	195.11	2.62	**2.96**	2.53	2.77	2.26

TABLE III
THE PERFORMANCE OF VARIOUS MODELS WHEN LEARNING DATASETS WITH IMAGES OF HIGHER COMPLEXITY.

	SSIM				PSNR			
Datasets	LGM [23]	LD-GAN	BE-Stu [49]	LTS [17]	LGM [23]	LD-GAN	BE-Stu [49]	LTS [17]
CelebA	0.05	0.57	0.56	0.25	12.06	19.07	19.26	13.33
CACD	0.01	0.61	0.46	0.33	10.53	18.92	17.06	14.52
CIFAR10	0.06	0.46	0.25	0.24	12.85	17.19	16.25	13.25
Sub-ImageNet	0.06	0.46	0.26	0.24	12.65	17.16	16.25	13.14
SVHN	0.20	0.66	0.50	0.57	13.54	13.65	13.08	12.75
MNIST	0.89	0.90	0.89	0.89	22.02	21.61	21.18	21.47
Average	0.21	**0.61**	0.49	0.42	13.94	**17.93**	17.18	14.74

LD-GAN and LTS outperform the VAE-based models such as LGM on all past tasks except MNIST, a result which is explained by Theorem 3. GANs usually produces better generative replay samples than VAE-based models and thus can reduce the Wasserstein distance term in RHS of Eq. (25), resulting in better performance. Since RMNIST shares similar visual concepts with MNIST, and thus none of the models suffers from forgetting when considering MNIST. We also evaluate the performance of various models in terms of Fréchet Inception Difference (FID) and Inception Score (IS) criteria under the MSFIR learning setting. The results are provided in Table II. Furthermore, the number of parameters of various models is provided in Table IV. From Tables II and IV we can observe that the proposed LD-GAN uses fewer parameters and achieves better performance than other models.

In the following, we consider randomly collecting 60,000 and 10,000 samples, from ImageNet database [56], as the training and testing set, respectively, creating the sub-ImageNet database. Then we consider a sequence of six tasks including CelebA, CACD, CIFAR10, Sub-ImageNet, SVHN and MNIST, namely CCCSSM. In Table III we evaluate LD-GANs, on the CCCSSM, which is more challenging than MSFIR, because of the complexity of the images from CCCSSM. These results show that the proposed LD-GANs outperforms other methods under this challenging learning setting. Finally, we present the visual results in Fig. 2 where the Teacher and Student alternatively are used to generate and reconstruct images. These results show that the proposed LD-GANs can provide better image generations and reconstructions when compared to LTS [17].

TABLE IV
THE NUMBER OF PARAMETERS NEEDED BY VARIOUS MODELS FOR LEARNING MSFIR.

LGM [23]	LD-GAN	BE-Stu [49]	LTS [17]	LIMix-Stu [2]
6.5×10^7	3.3×10^7	2.1×10^8	4.1×10^7	1.8×10^8

C. Lifelong learning evaluation on natural images

We create two subsets, from ImageNet database [56], contain a wide diversity of natural images, by randomly selecting 30,000 images from the original ImageNet for each subset, called Sub1 and Sub2, respectively. We should emphasize that we do not choose a subset of classes, but random images from all classes. This means that the chosen 60,000 selected images can cover all classes in various proportions. In addition, we ensure that there is no single image to be present in both Sub1 and Sub2. Then we train all models under CIFAR10, Sub1 and Sub2, by using a learning rate of 0,0002 and considering 20 epochs for each task training. The IS results when reconstructing 5,000 testing images, are provided in Table V. We evaluate the generated images quality by considering the FID score after each task switch, with the results reported in Table VI. The proposed LD-GANs framework outperforms other generative replay approaches, such as CURL [48] or LGM [23], by a large margin in terms of both IS and FID scores. The visual results of the reconstructed and generated images, after the lifelong

TABLE V
IS SCORE EVALUATED AFTER THE LIFELONG LEARNING OF CIFAR10 AND SUB-IMAGENET DATABASES SUB1 AND SUB2.

Dataset	LD-GANs	CURL [48]	LGM [23]
CIFAR10	**4.58**	3.46	3.41
Sub1	**4.69**	3.64	3.28
Sub2	**4.73**	3.63	3.32

TABLE VI
FID SCORE EVALUATED AFTER THE LIFELONG LEARNING OF CIFAR10 AND SUB-IMAGENET DATABASES SUB1 AND SUB2.

Tasks	LD-GANs	CURL [48]
First task	**62.85**	155.59
Second task	**59.27**	166.47
Third task	**60.35**	169.28

learning when considering a long sequence of tasks such as CCCSSM, where the first 'S' represents Sub1 and Sub2 data, are presented in Fig. 3-(c) and (e), respectively. We can observe that the LD-GANs can give higher-quality generations as well as better image reconstructions than CURL, whose reconstructions and generations are shown in Fig. 3-(b) and (d), respectively.

D. Learning interpretable representations across domains

The Student module of the lifelong LD-GANs Teacher-Student network, described in Section III-D, is trained considering the ELBO criterion from Eq. (10), under CelebA [53] to CACD [54], and CelebA to 3D-Chair [59] lifelong learning, respectively. Two latent vectors encoding distinct images are interpolated, and the results are shown in Fig. 4. These results indicate that a given image can be progressively transformed into another one, which is completely different from the initial image, even when these two images are sourced from completely different domains. This demonstrates that the higher-quality knowledge transferring process between the Teacher and Student can allow a simply designed latent variable model, from the Student's representation, to capture over time both continuous and domain-specific features.

We also test the disentanglement ability by training the LD-GANs Teacher-Student using CelebA and 3D-Chairs databases considering the loss defined by Eq. (11), where we consider $\beta = 4$. We then change a latent variable while fixing the others and the results are shown in Fig. 5 for CelebA to 3D-Chair lifelong learning. In Fig. 5-(a) and (b) we show how we can change face orientation and width, respectively, while in Fig 5-(c) and (d), we change the back side of a chair and chair's size, respectively.

E. Supervised learning

In this section, we focus on the lifelong classification task. We consider the classification accuracy as the performance criterion, which was also used for testing other continual supervised learning methods [2], [60]. We consider three datasets, MNIST [50] and SVHN [51] and Fashion [52] and resize their images to $32 \times 32 \times 3$ pixels. We consider a simple CNN consisting of two convolution layers for both the decoder and encoder of the Student module as well as for each expert. The number of training epochs for each task is set to 10. We then evaluate the performance of various models under supervised learning and report the results in Tables VII and VIII, respectively, where "M-S" represents the model that firstly learns MNIST and afterwards SVHN, while "M-F" represents that case when reversing the order of training for the two databases. These results show that the proposed LD-GANs outperforms other supervised lifelong learning baselines.

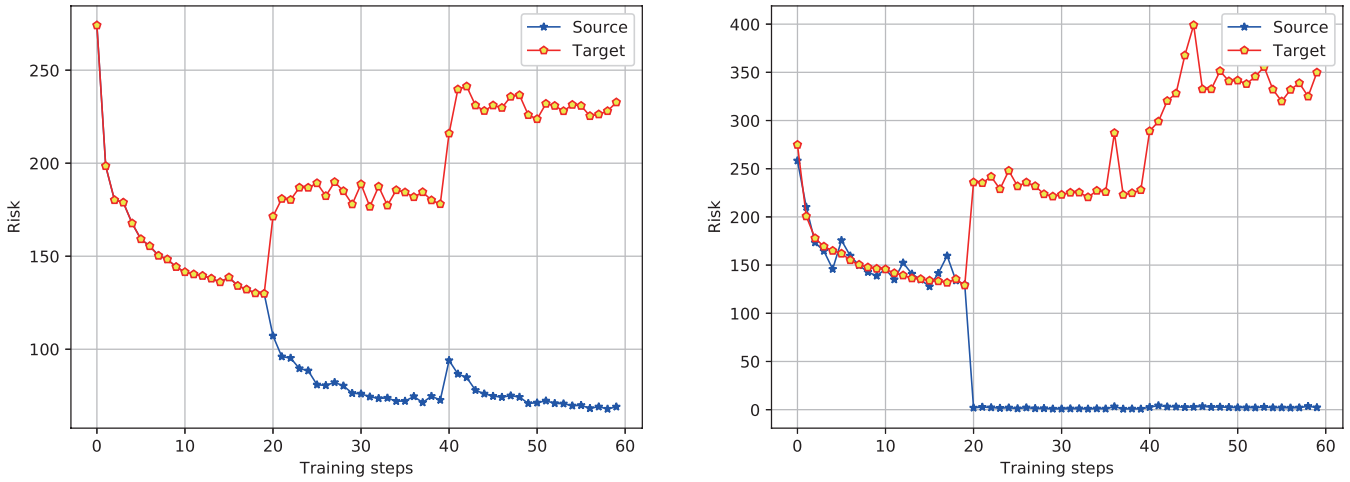
TABLE VII
CLASSIFICATION RESULTS FOLLOWING THE LIFELONG LEARNING OF MNIST AND SVHN DATABASES. THE RESULTS FROM OTHER BASELINES ARE CITED FROM [17].

Methods	Testing data set	Lifelong	Accuracy
LTS [17]	MNIST	M-S	96.66
MemoryGANs [57]	MNIST	M-S	96.04
LGM [23]	MNIST	M-S	96.59
LD-GANs	MNIST	M-S	**96.89**
LTS [17]	SVHN	M-S	80.15
MemoryGANs [57]	SVHN	M-S	80.03
LGM [23]	SVHN	M-S	80.77
LD-GANs	SVHN	M-S	**81.02**
LTS [17]	MNIST	S-M	98.80
MemoryGANs [57]	MNIST	S-M	98.29
LGM [23]	MNIST	S-M	98.56
LD-GANs	MNIST	S-M	**98.93**
LTS [17]	SVHN	S-M	80.39
MemoryGANs [57]	SVHN	S-M	79.34
LGM [23]	SVHN	S-M	76.76
LD-GANs	SVHN	S-M	**80.45**

In the following, we also investigate the performance of the proposed approach in a more challenging setting, the learning of a long sequence of tasks. We train the LD-GANs with the images from the database sequence called MSFIIC, containing MNIST, SVHN, Fashion, InverseFashion, InverseMNIST and CIFAR10. We also create InverseFashion and InverseMNIST, by inversing each pixel in all images from Fashion and MNIST databases, respectively. We report the results in Table IX, which indicate that the GAN-based models achieve better results for each task than VAE-based methods.

F. Empirical results for forgetting analysis

In this section, we investigate the forgetting behaviour of the proposed LD-GANs model. Firstly, for evaluating how the error is accumulated when learning each new task, we train LD-GANs considering CIFAR10, CACD and MNIST (CCM) databases under the lifelong learning setting. We also train a task-inference model that returns the task label for each sample, which is used to select the generated samples that belong to CIFAR10. We then evaluate the target risk (CIFAR10) and source risk (selected generated dataset) in each



(a) Risks on CIFAR10.

(b) Risks on all previously learnt tasks.

Fig. 6. The risk estimated by LD-GANs under CIFAR10, CACD and MNIST lifelong learning, where 20 training epochs are considered for learning each database.

TABLE VIII
CLASSIFICATION RESULTS FOLLOWING THE LIFELONG LEARNING OF MNIST AND FASHION. THE RESULTS FROM OTHER BASELINES ARE CITED FROM [17].

Methods	Testing data set	Lifelong	Accuracy
LTS [17]	MNIST	M-F	98.51
LGM [23]	MNIST	M-F	97.29
MemoryGANs [57]	MNIST	M-F	98.15
LD-GANs	MNIST	M-F	**98.62**
LTS [17]	Fashion	M-F	91.49
LGM [23]	Fashion	M-F	91.71
MemoryGANs [57]	Fashion	M-F	91.35
LD-GANs	Fashion	M-F	**91.68**
LTS [17]	MNIST	F-M	98.42
LGM [23]	MNIST	F-M	98.85
MemoryGANs [57]	MNIST	F-M	98.52
LD-GANs	MNIST	F-M	**98.95**
LTS [17]	Fashion	F-M	89.35
LGM [23]	Fashion	F-M	86.05
MemoryGANs [57]	Fashion	F-M	89.13
LD-GANs	Fashion	F-M	**89.44**

TABLE IX
CLASSIFICATION RESULTS UNDER MSFIIC LIFELONG LEARNING.

Dataset	LD-GAN	LGM [23]	MeRGANs [20]
MNIST	83.33	81.08	85.87
SVHN	53.88	24.28	32.85
Fashion	76.05	49.70	61.75
InverseFashion	77.76	38.39	64.38
InverseMNIST	96.96	80.86	95.52
CIFAR10	54.44	56.79	58.71
Average	**73.74**	55.18	66.51

training epoch and report the results in Fig 6-(a). From this plot, we observe that the target risk is gradually increasing while the source risk is rather constant, when learning other tasks after the initial CIFAR10 database. This result indicates that the lower source risk can not guarantee a good generalization performance due to the accumulated errors, as discussed in Lemma 2. Additionally, we also evaluate the risk on all previously learnt tasks when learning each task and present the results in Fig 6-(b), which shows that the model tends to have degenerated performance on all previously learnt tasks. This is due to the accumulated errors caused by the GRM process during each task learning, according to Theorem 3.

VI. CONCLUSION

In this research study we propose a novel lifelong generative learning model called the Lifelong Dual Generative Adversarial Nets (LD-GANs), which is used to learn successively multiple tasks. In order to train the LD-GANs, we propose the Lifelong Adversarial Knowledge Distillation (LSKD), representing an end-to-end memory efficient method for accumulating information from several tasks. LD-GANs consists of two GAN generators, implementing the Teacher and Assistant, which under LSKD are used to generate data by teaching each other, each time when receiving a new task for learning. This model is extended to a Teacher-Student framework in order to learn data representations over time. Based on the higher-quality knowledge transfer from LD-GANs to the Student model, the network can capture shared and task-specific parameters across tasks over time. We also introduce a new theoretical framework based on the Wasserstein distance, which provides new insights into the forgetting behaviour of the Student. From both theoretical concepts and through extensive experimental results we show that the proposed methodology is better than other lifelong learning approaches.

REFERENCES

[1] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[2] F. Ye and A. G. Bors, "Lifelong infinite mixture model based on knowledge-driven Dirichlet process," in *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 695–10 704.

[3] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proc. of the National Academy of Sciences (PNAS)*, vol. 114, no. 13, pp. 3521–3526, 2017.

[4] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[5] F. Ye and A. G. Bors, "Lifelong mixture of variational autoencoders," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 461–474, 2023.

[6] G. Sun, C. Yang, J. Liu, L. Liu, X. Xu, and H. Yu, "Lifelong metric learning," *IEEE Trans. on Cybernetics*, vol. 49, no. 8, pp. 3168–3179, 2019.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 2014, pp. 2672–2680.

[8] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2020, pp. 8107–8116.

[9] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8789–8797.

[10] H. Zhu, X. Peng, V. Chandrasekhar, L. Li, and J.-H. Lim, "DehazeGAN: When image dehazing meets differential programming." in *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2018, pp. 1234–1240.

[11] B. Li, Y. Gou, S. Gu, J. Z. Liu, J. T. Zhou, and X. Peng, "You only look yourself: Unsupervised and untrained single image dehazing neural network," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1754–1767, 2021.

[12] F. Ye and A. G. Bors, "InfoVAEGAN: Learning joint interpretable representations by information maximization and maximum likelihood," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, 2021, pp. 749–753.

[13] ——, "Learning joint latent representations based on information maximization," *Information Sciences*, vol. 567, pp. 216–236, 2021.

[14] ——, "Learning latent representations across multiple data domains using lifelong VAEGAN," in *Proc. European Conf on Computer Vision (ECCV), vol. LNCS 12365*, 2020, pp. 777–795.

[15] A. Seff, A. Beatson, D. Suo, and H. Liu, "Continual learning in generative adversarial nets," *arXiv preprint arXiv:1705.08395*, 2017.

[16] F. Ye and A. G. Bors, "Lifelong twin generative adversarial networks," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, 2021, pp. 1289–1293.

[17] ——, "Lifelong teacher-student network learning," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6270–6296, 2022.

[18] ——, "Lifelong learning of interpretable image representations," in *Proc. Int. Conf. on Image Processing Theory, Tools and Applications (IPTA)*, 2020.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[20] C. Wu, L. Herranz, X. Liu, J. van de Weijer, and B. Raducanu, "Memory replay GANs: Learning to generate new categories without forgetting," in *Proc. Advances In Neural Inf. Proc. Systems (NIPS)*, 2018, pp. 5962–5972.

[21] A. Achille, T. Eccles, L. Matthey, C. Burgess, N. Watters, A. Lerchner, and I. Higgins, "Life-long disentangled representation learning with cross-domain latent homologies," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 2018, pp. 9873–9883.

[22] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2017.

[23] J. Ramapuram, M. Gregorova, and A. Kalousis, "Lifelong generative modeling," *Neurocomputing*, vol. 404, pp. 381–400, 2020.

[24] A. Kuzina, E. Egorov, and E. Burnaev, "BooVAE: Boosting approach for continual learning of VAE," in *Proc. Advances in Neural Inf. Proc. Systems (NeurIPS)*, 2021, pp. 17 889–17 901.

[25] F. Ye and A. G. Bors, "Deep mixture generative autoencoders," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. :5789–5803, 2021.

[26] ——, "Mixtures of variational autoencoders," in *Proc. Int. Conf. on Image Processing Theory, Tools and Applications (IPTA)*, 2020, pp. 1–6.

[27] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner, "Variational continual learning," in *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1710.10628*, 2017.

[28] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, "Progress & compress: A scalable framework for continual learning," in *Proc. of International Conference on Machine Learning (ICML), vol. 80*, 2018, pp. 4535–4544.

[29] J. Martens and R. B. Grosse, "Optimizing neural networks with Kronecker-factored approximate curvature," in *Proc. of the International Conference on Machine Learning*, 2015, pp. 2408–2417.

[30] H. Ahn, S. Cha, D. Lee, and T. Moon, "Uncertainty-based continual learning with adaptive regularization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 4394–4404.

[31] W. Chen, B. Chen, Y. Liu, X. Cao, A. Zhao, H. Zhang, and L. Tian, "Max-margin deep diverse latent Dirichlet allocation with continual learning," *IEEE Trans. on Cybernetics*, vol. 52, no. 7, pp. 5639 – 5653, 2022.

[32] J. Le, X. Lei, N. Mu, H. Zhang, K. Zeng, and X. Liao, "Federated continuous learning with broad network architecture," *IEEE Trans. on Cybernetics*, vol. 51, no. 8, pp. 3874–3888, 2021.

[33] M. Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 2016, pp. 469–477.

[34] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 2849–2857.

[35] M. Gong, Y. Xu, C. Li, K. Zhang, and K. Batmanghelich, "Twin auxiliary classifiers GAN," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 1328–1337.

[36] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 70*, 2017, pp. 214–223.

[37] ——, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.

[38] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 2017, pp. 5769–5779.

[39] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Advances in Neural Information Processing Systems*, 2017, pp. 742–751.

[40] M. Phuong and C. Lampert, "Towards understanding knowledge distillation," in *Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 97*, 2019, pp. 5142–5151.

[41] M. G. Andrey Malinin, Bruno Mlodozeniec, "Ensemble distribution distillation," in *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1905.00076*, 2020.

[42] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 80*, 2018, pp. 1607–1616.

[43] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," in *Proc. Advances in Neural Inf. Proc. Systems (NIPS)*, 2017, pp. 2990–2999.

[44] G. Desjardins, A. Courville, and Y. Bengio, "Disentangling factors of variation via generative entangling," *arXiv preprint arXiv:1210.5474*, 2012.

[45] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, 2018, pp. 2615–2625.

[46] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," in *Proc. NIPS Workshop on Learning Disentangled Representation, arXiv preprint arXiv:1804.03599*, 2017.

[47] I. Redko, A. Habrard, and M. Sebban, "Theoretical analysis of domain adaptation with optimal transport," in *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), vol. LNCS 10535*, 2017, pp. 737–753.

[48] D. Rao, F. Visin, A. A. Rusu, Y. W. Teh, R. Pascanu, and R. Hadsell, "Continual unsupervised representation learning," in *Advances Neural Information Processing Systems (NIPS)*, 2019, pp. 7647–7657.

[49] Y. Wen, D. Tran, and J. Ba, "BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning," in *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:2002.06715*, 2020.
[50] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[51] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
[52] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.
[53] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2015, pp. 3730–3738.
[54] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. European Conf on Computer Vision (ECCV), vol. LNCS 8694*, 2014, pp. 768–783.
[55] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. of Toronto, Tech. Rep., 2009.
[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[57] C. Wu, L. Herranz, X. Liu, Y. Wang, J. van de Weijer, and B. Raducanu, "Memory replay GANs: Learning to generate new categories without forgetting," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 2018, pp. 5962–5972.
[58] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, 2010, pp. 2366–2369.
[59] M. Aubry, D. Maturana, A. A. Efros, B. Russell, and J. Sivic, "Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3762–3769.
[60] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, D. Dokania, P. H. S. Torr, and M. Ranzato, "On tiny episodic memories in continual learning," *arXiv preprint arXiv:1902.10486*, 2019.

Fei Ye is currently a PHD candidate in computer science from the University of York. He received the bachelor degree from Chengdu University of Technology, China, in 2014 and the master degree in computer science and technology from Southwest Jiaotong University, China, in 2018. His research topics includes deep generative image models, lifelong learning and mixture models.

Adrian G. Bors (Senior Member, IEEE) received the M.Sc. degree in electronics engineering from the Polytechnic University of Bucharest, Bucharest, Romania, in 1992, and the Ph.D. degree in informatics from the University of Thessaloniki, Thessaloniki, Greece, in 1999. In 1999 he joined the Department of Computer Science, Univ. of York, U.K., where he is currently an Associate Professor. Dr. Bors was a Research Scientist at Tampere Univ. of Technology, Finland, a Visiting Scholar at the Univ. of California at San Diego (UCSD), and an Invited Professor at the Univ. of Montpellier, France. Dr. Bors has authored and co-authored more than 160 research papers, including 40 in journals. His research interests include computational intelligence, computer vision, pattern recognition and image processing.

Dr. Bors was a member of the organizing committees for IEEE WIFS 2021, IPTA 2020, IEEE ICIP 2018, BMVC 2016, IPTA 2014, CAIP 2013, and IEEE ICIP 2001. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2010 to 2014 and the IEEE TRANSACTIONS ON NEURAL NETWORKS from 2001 to 2009. He was a Co-Guest Editor for a special issue on Machine Vision for the International Journal for Computer Vision in 2018 and the Journal of Pattern Recognition in 2015.