White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/168546/

Version: Accepted Version

Proceedings Paper:
Ye, Fei and Bors, Adrian Gheorghe (2020) Learning latent representations across multiple data domains using Lifelong VAEGAN. In: European Conference in Computer Vision (ECCV). Lecture Notes in Computer Science. Springer, Manchester, UK, pp. 777-795.

https://doi.org/10.1007/978-3-030-58565-5_46

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Learning latent representations across multiple data domains using Lifelong VAEGAN

Fei Ye and Adrian G. Bors

Department of Computer Science, University of York, York YO10 5GH, UK

Abstract. The problem of catastrophic forgetting occurs in deep learning models trained on multiple databases in a sequential manner. Recently, generative replay mechanisms (GRM), have been proposed to reproduce previously learned knowledge aiming to reduce the forgetting. However, such approaches lack an appropriate inference model and therefore can not provide latent representations of data. In this paper, we propose a novel lifelong learning approach, namely the Lifelong VAEGAN (L-VAEGAN), which not only induces a powerful generative replay network but also learns meaningful latent representations, benefiting representation learning. L-VAEGAN can allow to automatically embed the information associated with different domains into several clusters in the latent space, while also capturing semantically meaningful shared latent variables, across different data domains. The proposed model supports many downstream tasks that traditional generative replay methods can not, including interpolation and inference across different data domains.

Keywords: Lifelong learning, Representation learning, Generative modeling, VAEGAN model.

1 Introduction

The lifelong learning framework describes an intelligent learning process capable of remembering all previously learned knowledge from multiple sources, such as different databases [47]. The ability of continuous, or lifelong learning, is an inherent characteristic of humans and animals, which helps them to adapt to the environment during their entire life. However, such characteristics remain an open challenge for deep learning models. The current state-of-the-art deep learning approaches perform well on many individual databases [18, 55], but suffer from catastrophic forgetting when attempting to learn data associated with new tasks [3, 13, 19, 51, 62, 65]. For example when a deep neural network is trained on a new database, its parameters are updated in order to learn new information while their previous values are lost. Consequently, their performance on the previously learnt tasks degenerates.

In order to alleviate the catastrophic forgetting problem, memory-based approaches use a buffer to store a small subset of previously seen data samples [2, 4, 8]. However, such approaches cannot be seen as lifelong learning models and

they do not scale well when increasing the number of databases defining different tasks. Shin *et al.* [60] proposed a learning model employing the Generative Replay Mechanism (GRM). The idea of the GRM is to train a generative replay network to reproduce previously learnt knowledge by using adversarial learning. A classifier is then trained using jointly generative replay data and data sampled from the current database. This approach was only applied on prediction tasks. Although recent studies, such as the one from [64], have been used to generate images from new classes without forgetting, generative replay approaches do not learn the representation of data and therefore can not be extended to be used in a broad range of tasks. Learning meaningful and disentangled representations of data was shown to benefit many tasks, but they have not been explored so far within the lifelong learning methodology [1, 52]. In this paper, we propose a new lifelong learning model which not only learns a GRM but also induces accurate inference models, benefiting on representation learning.

This research study brings the following contributions :

1) In order to address the drawbacks of generative replay approaches, we propose a novel lifelong learning model aiming to learn informative latent variables over time.
2) We show that the proposed lifelong learning model can be extended for unsupervised, semi-supervised and supervised learning with few modifications.
3) We propose a two-step optimization algorithm to train the proposed model. The latent representation learned by the proposed model can capture both task-specific generative factors and semantic meaningful shared latent variables across different domains over time.
4) We provide a theoretical insight into how GRM models are used for lifelong learning in artificial systems.

2 Related works

The lifelong learning was approached in previous research studies from three different perspectives: by using regularization, dynamic architectures, and by employing memory replay. Regularization approaches in order to alleviate catastrophic forgetting add an auxiliary term that penalizes changes in the weights when the model is trained on a new task [14, 26, 29, 33, 38, 48, 54, 60]. Dynamic architectures would increase the number of neurons and network layers in order to adapt to learning new information [58]. Most memory replay approaches are using generative models, such as either Generative Adversarial Networks (GANs) [21] or Variational Autoencoders (VAEs) [32]) to replay the previously learnt knowledge. For instance, Wu *et al.* [64] proposed a novel lifelong generative framework, namely Memory Replay GANs (MeRGANs), which mainly generates images from new categories under the lifelong learning setting. The Lifelong GAN [67] employs image to image translation. However, both models from [64, 67] lack an image inference procedure and Lifelong GAN would need to load all previously learnt data for the generation task. Approaches employing both generative and inference mechanisms are based on the VAE framework [1,

50]. However, these approaches have degenerating performance when learning high-dimensional data, due to lacking a powerful generator.

Hybrid VAE-GAN methods learn an inference model from a GAN model, which can also capture data representations, which is specific to the VAE. Adversarial learning is performed in order to match either the data distribution [36], the latent variables distribution [42], or their joint distributions [10, 15, 16, 37, 43, 49, 61]. These methods perform well only when trained on a single dataset and their performance would degenerate when learning a new task.

This paper is the first research study to propose a novel hybrid lifelong learning model, which not only addresses the drawback of the existing hybrid methods but also provides inference mechanisms for the GRM, benefiting on many downstream tasks across domains under the lifelong learning framework. The approach proposed in this paper also addresses disentangled representation learning [35] in the context of lifelong learning. Many recent approaches would aim to modify the VAE framework in order to learn a meaningful representation of the data by imposing a large penalty on the Kullback-Leibler (KL) divergence term [7, 17, 25] or on the total correlation latent variables [12, 20, 28, 30]. These approaches perform well on independent and identically distributed data samples from a single domain. However, they are unable to learn the information from piecewise changing stationary data from multiple databases, because they suffer from catastrophic forgetting.

3 The Lifelong VAEGAN

In this section, we introduce the optimization algorithm used for training the proposed model when learning several databases without forgetting.

3.1 Problem formulation

The lifelong learning problem consists of learning a sequence of of K tasks, each characterized by a distinct database, corresponding to the data distributions $p(\mathbf{x}^1), p(\mathbf{x}^2), \ldots, p(\mathbf{x}^K)$. During the k-th database learning, we only access the images sampled from $p(\mathbf{x}^k)$. Most existing lifelong learning approaches focus on prediction or regression tasks. Meanwhile, in this research study we focus on modelling the overall data distribution $p(\mathbf{x})$, by learning latent representations over time:

$$p(\mathbf{x}) := \int \prod_{i=1}^{K} p(\mathbf{x}^i|\mathbf{z})p(\mathbf{z})d\mathbf{z} \tag{1}$$

where $\mathbf{z}$ represents the latent variables defining the information of all previously learnt databases. Besides addressing unsupervised learning, we also incorporate discrete variables into the optimization path in order to capture category discriminating information. Moreover, we also consider the semi-supervised learning problem where we consider that in each dataset we only have some labelled data while the rest are unlabeled.

3.2 Data generation from prior distributions

In the following we aim to learn two separate latent representations for capturing discrete and continuous variations of data. The discrete data are denoted as $\mathbf{c} = \{\mathbf{c}_i | i = 1, \ldots, L\}$ where L is the dimension of the discrete variable space, while the continuous variables $\mathbf{z}$ are sampled from a normal distribution $\mathcal{N}(0, \mathrm{I})$. We also consider the domain variable $\mathbf{a} = \{\mathbf{a}_j | j = 1, \ldots, K\}$, defining each database and aiming to capture the information characterizing its task. The generation process considering the three latent variables $\mathbf{c}$, $\mathbf{z}$ and $\mathbf{a}$ is defined as:

$$\begin{aligned} &\mathbf{c} \sim \mathrm{Cat}\left(K = L, p = 1/K\right), \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}), \\ &\mathbf{a} \sim \mathrm{Cat}\left(p_1, p_2, \ldots, p_K\right), \mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{a}, \mathbf{c}), \end{aligned} \tag{2}$$

where $\mathrm{Cat}(\cdot)$ is the Categorical distribution, and $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{a}, \mathbf{c})$ is the distribution characterizing the generator implemented by a neural network with trainable parameters, θ. By incorporating the domain variables $\mathbf{a}$ in the inference model helps to generate images characteristic to a specific task. We consider the Wasserstein GAN (WGAN) [5] loss with the gradient penalty [22], which is defined by :

$$\begin{aligned} \min_G \max_D \mathcal{L}^G_{GAN}(\theta, \omega) =& \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}), \mathbf{c}\sim p(\mathbf{c}), \mathbf{a}\sim p(\mathbf{a})}[D(G(\mathbf{c}, \mathbf{z}, \mathbf{a}))] - \\ & \mathbb{E}_{p(\mathbf{x})}[D(\mathbf{x})] + \lambda \mathbb{E}_{\tilde{\mathbf{x}}\sim p(\tilde{\mathbf{x}})}[(\|\nabla_{\tilde{\mathbf{x}}} D(\tilde{\mathbf{x}})\|_2 - 1)^2] \end{aligned} \tag{3}$$

where we introduce a Discriminator D, defined by the trainable parameters ω, $p(\mathbf{x})$ denotes the true data distribution, and the third term is the gradient weighted by the penalty λ. The adversarial loss allows the Generator and Discriminator to be trained alternately such that the Discriminator aims to distinguish real from generated data, while the Generator tends to fool the Discriminator through aiming to generate realistic data [5, 21].

3.3 Training the inference model

Most GAN-based lifelong methods [60, 64, 66] do not learn an accurate inference model and therefore can not derive a meaningful data representation. For the model proposed in this paper, we consider three differentiable non-linear functions $f_\varsigma(\cdot)$, $f_\varepsilon(\cdot)$, $f_\delta(\cdot)$, aiming to infer three different types of latent variables $\{\mathbf{z}, \mathbf{c}, \mathbf{a}\}$. We implement $f_\varsigma(\cdot)$ by using the Gaussian distribution $\mathcal{N}(\mu, \sigma)$ where $\mu = \mu_\varsigma(\mathbf{x})$ and $\sigma = \sigma_\varsigma(\mathbf{x})$ are given by the outputs of a neural network with trainable parameters ς. We use the reparameterization trick [32, 56] for sampling $\mathbf{z} = \mu + \pi \otimes \sigma$, where π is a random noise vector sampled from $\mathcal{N}(0, \mathrm{I})$, in order to ensure end-to-end training.

Discrete variables. We can not sample the discrete latent variables $\mathbf{a}$ and $\mathbf{c}$ from $f_\varepsilon(\cdot)$ and $f_\delta(\cdot)$, respectively, because the categorical representations are non-differentiable. In order to mitigate this, we use the Gumbel-Max trick [23, 41] for achieving the differentiable relaxation of discrete random variables. The Gumbel-softmax trick was also used in [17, 27, 40, 63] and its capability of reducing the variation of gradients was shown in [63].

The sampling process of discrete latent variables is defined as:

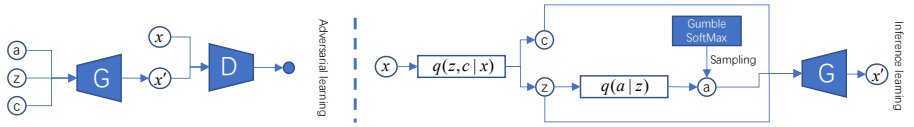
$$\mathbf{a}_j = \frac{\exp((\log \mathbf{a}'_j + \mathbf{g}_j)/T)}{\sum_{i=1}^{K} \exp((\log \mathbf{a}'_i + \mathbf{g}_i)/T)} \tag{4}$$

where $\mathbf{a}'_i$ is the i-th entry of the probability defined by the softmax layer characterizing $f_\varepsilon(\cdot)$ and $\mathbf{a}_j$ is the continuous relaxation of the domain variable, while $\mathbf{g}_k$ is sampled from the distribution Gumbel$(0, 1)$ and T is the temperature parameter that controls the degree of smoothness. We use the Gumbel softmax trick for sampling both the domain $\mathbf{a}$ and the discrete $\mathbf{c}$ variables.

The log-likelihood objective function. GANs lack an inference mechanism, preventing them to capture data representations properly. In this paper we propose to maximize the sample log-likelihood for learning the inference models, defined by $p(\mathbf{x}) = \int\int\int p(\mathbf{x}|\mathbf{z}, \mathbf{a}, \mathbf{c})p(\mathbf{z}, \mathbf{a}, \mathbf{c})\, d\mathbf{z}\, d\mathbf{a}\, d\mathbf{c}$, which is intractable in practice. We therefore derive the following lower bound on the log-likelihood, which is characteristic to VAEs, by introducing variational distributions:

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(\theta, \varsigma, \varepsilon, \delta) =& \mathbb{E}_{q_{\varsigma,\varepsilon,\delta}(\mathbf{z},\mathbf{a},\mathbf{c}|\mathbf{x})} \log[p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{a}, \mathbf{c})] - D_{KL}[q_\varsigma(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \\ &- \mathbb{E}_{q_\varsigma(\mathbf{z}|\mathbf{x})} D_{KL}[q_\varepsilon(\mathbf{a}|\mathbf{x})||p(\mathbf{a}|\mathbf{z})] - D_{KL}[q_\delta(\mathbf{c}|\mathbf{x})||p(\mathbf{c})]\end{aligned} \tag{5}$$

where $q_\varsigma(\mathbf{z}|\mathbf{x})$, $q_\varepsilon(\mathbf{a}|\mathbf{z})$, $q_\delta(\mathbf{c}|\mathbf{x})$ are variational distributions modelled by $f_\varsigma(\cdot)$, $f_\varepsilon(\cdot)$, $f_\delta(\cdot)$, respectively. For the third term from (5), we sample from the empirical distribution and then sample $\mathbf{z}$ from $q_\delta(\mathbf{c}|\mathbf{x})$. $p(\mathbf{a}|\mathbf{z})$ is the prior distribution Cat$(p_1, \ldots, p_K)$, where p_i denotes the probability of the sample belonging to the i-th domain. We consider $q_\varepsilon(\mathbf{a}|\mathbf{z})$ as the task-inference model which aims to infer the task ID for the given data samples.



(a) Generator network (b) Inference network

Fig. 1. The graph structure for the proposed Lifelong VAEGAN (L-VAEGAN) model, where G and D denote the Generator and Discriminator, respectively.

For the supervised learning setting, the auxiliary information such as the class labels can be used to guide the inference model. We minimise the cross-entropy loss $\eta(\cdot, \cdot)$ for $q_\delta(\mathbf{c}|\mathbf{x})$ and $q_\varepsilon(\mathbf{a}|\mathbf{z})$ as:

$$\mathcal{L}_{\mathbf{a}}(\varepsilon) = \mathbb{E}_{(\mathbf{x},\mathbf{a}^*)\sim(\mathbf{X},\mathbf{A}),\mathbf{z}\sim q_\varsigma(\mathbf{z}|\mathbf{x})} \eta(q_\varepsilon(\mathbf{a}|\mathbf{z}), \mathbf{a}^*) \tag{6}$$

$$\mathcal{L}_{\mathbf{c}}(\delta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathbf{X},\mathbf{Y})} \eta(q_\delta(\mathbf{c}|\mathbf{x}), \mathbf{y}) \tag{7}$$

where $\mathbf{X}$ and $\mathbf{Y}$ represent the empirical data and target distributions, respectively. $\mathbf{a}^*$ is the variable drawn from $\mathbf{A}$ which represents the Categorical distribution Cat$(m_1, \ldots, m_k)$, where m_i is the probability of seeing i-th task, characterizing the corresponding database. The graph structure of the Generator and Inference networks of the proposed Lifelong VAEGAN (L-VAEGAN) is shown in Figs. 1-a and 1-b, respectively, where the variable $\mathbf{a}$ is conditioned on $\mathbf{z}$. The

proposed model is flexible to be extended for recognizing new tasks by automatically appending the domain variable $\mathbf{a}$ and optimizing the task-inference model $q_\varepsilon(\mathbf{a}|\mathbf{z})$ when faced with learning a new task.

4 Theoretical analysis of the GRM

In this section, we analyze the GRM used in lifelong learning.

Definition 1. *We define the distribution modelling the lifelong learned data as $p(\tilde{\mathbf{x}}^t)$, which is encoded through $G_{\gamma_t}(\mathbf{z}, \mathbf{c}, \mathbf{d})$. The assumption is that the network has learnt the information from all given databases $t = 1, \ldots, K$, and this information is stored, refined and processed across various tasks, where γ_t is the generator parameter updated after the t-th task learning.*

Definition 2. *Let us define*

$$p(\tilde{\mathbf{x}}^t|\tilde{\mathbf{x}}^{t-1}, \mathbf{x}^t) = \exp(-\Gamma(p(\tilde{\mathbf{x}}^{t-1}, \mathbf{x}^t), p(\tilde{\mathbf{x}}^t))) \tag{8}$$

as the probability of generated data $\tilde{\mathbf{x}}^t$ when observing $\tilde{\mathbf{x}}^{t-1}$ and $\mathbf{x}^t$, where $\Gamma(\cdot)$ is a probabilistic measure of comparison between two distributions, which can be the f-divergence [46], or the Wasserstein distance [5] (Earth-mover distance).

Theorem 1. *By marginalizing over $\tilde{\mathbf{x}}^{t-1}$ and $\mathbf{x}^t$, on $p(\tilde{\mathbf{x}}^t|\tilde{\mathbf{x}}^{t-1}, \mathbf{x}^t)$, the resulting marginal distribution $p(\tilde{\mathbf{x}}^t)$ encodes the statistical correlations from all previously learnt distributions.*

Proof. By using mathematical induction over the lifelong learning of the probabilities associated with various tasks, the marginal distribution is rewritten as:

$$p(\tilde{\mathbf{x}}^t) = \int \ldots \int p(\tilde{\mathbf{x}}^1) \prod_{i=0}^{t-2} p(\tilde{\mathbf{x}}^{t-i}|\tilde{\mathbf{x}}^{t-i-1}, \mathbf{x}^{t-i}) \prod_{i=0}^{t-2} p(\mathbf{x}^{t-i}) d\tilde{\mathbf{x}}^1 \ldots d\tilde{\mathbf{x}}^{t-1} d\mathbf{x}^2 \ldots d\mathbf{x}^t \tag{9}$$

Lemma 1. *The data probability $p(\tilde{\mathbf{x}}^t)$ approximates the true joint distribution $\prod_{i=1}^{t} p(\mathbf{x}^i)$ when all previously learnt distributions are the exact approximations to their target distributions while learning every given task.*

In the following we extend the theoretical analysis on the domain adaptation problem from [53] (Theorem 2) in order to analyze how the knowledge learned by GRMs is lost during the lifelong learning.

Theorem 2. *Let us consider two vector samples, one corresponding to the generated data $\{\nu_{t'} \in \mathrm{R}^{\mathrm{s}}|\nu_{t'} \sim p(\tilde{\mathbf{x}}^t)\}$ and another corresponding to the real data $\{\nu_t \in \mathrm{R}^{\mathrm{s}}|\nu_t \sim p(\mathbf{x}^t)\}$ of sizes n_t and $n_{t'}$. Then let $h^t(\cdot)$ be a new learned model trained on $\nu_{t'}$. For any $s' > s$ and $a' < \sqrt{2}$, there is a constant n_0 depending on s' satisfying that for any $\delta > 0$ and $\min(\nu_t, n_{t'}) \geq n_0 \max(\delta^{-(s'+2)}, 1)$. Then with the probability of at least $1 - \delta$ for all h^t, we have:*

$$E\left(h^t(\nu_t)\right) \leq E\left(h^t(\nu_{t'})\right) + W\left(\nu_t, \nu_{t'}\right) + \sqrt{2\log\left(\frac{1}{\delta}\right)/a'}\left(\sqrt{\frac{1}{n_t}} + \sqrt{\frac{1}{n_{t'}}}\right) + D \tag{10}$$

where $E(h^t(\nu_t)), E(h^t(\nu_{t'}))$ *denote the observed risk for* ν_t *and* $\nu_{t'}$*, respectively, and* $W(\nu_t, \nu_{t'})$ *is the Wassenstein distance between* ν_t *and* $\nu_{t'}$*. D is the combined error when we find the optimal model* $h^{t'} =: \arg\min_{h^t \in \mathcal{H}}(E(h^t(\nu_t)) + E(h^t(\nu_{t'}))$.

This theorem demonstrates that the performance of a model h^t degenerates on the empirical data distribution $p(\mathbf{x}^t)$. From Theorem 2, we conclude that the lifelong learning becomes a special domain adaptation problem in which the target and source domain are empirical data distributions from the current task and the approximation distribution $G_{\gamma_t}(\mathbf{z}, \mathbf{c}, \mathbf{d})$.

Lemma 2. *From Theorem 2 we have a bound on the accumulated errors across tasks during the lifelong learning.*

$$\begin{aligned}&\sum_{i=1}^{K} E\left(h^K(\nu_i)\right) \leq \sum_{i=1}^{K} E\left(h^K(\nu_{i^{(K)}})\right) + \\ &\mathbf{W}\left(\nu_i, \nu_{i^{(K)}}\right) + \sqrt{2\log\left(\frac{1}{\delta}\right)/a'}\left(\sqrt{\frac{1}{n_i}} + \sqrt{\frac{1}{n_{i^{(K)}}}}\right) + D_{(i^{(K-1)}, i^{(K)})},\end{aligned} \tag{11}$$

where $E\left(h^K(\nu_{i^{(K)}})\right)$ *denotes the observed risks on the probability measure* $\nu_{i^{(K)}}$ *formed by samples drawn from* $p(\tilde{\mathbf{x}}^i)$*, after they have been learned across K tasks.* $D_{(i^{(K-1)}, i^{(K)})}$ *is the combined error of an optimal model*

$$h^* = \arg\min_{h \in \mathcal{H}}(E\left(h^K(\nu_{i^{(K-1)}})\right) + E\left(h^K(\nu_{i^{(K)}})\right)) \tag{12}$$

Theorem 3. *By having a learning system, acquiring the information from the given databases, we can define* $\log p_\theta(\mathbf{x}^t, \tilde{\mathbf{x}}^{t-1})$ *as the joint model log-likelihood and* $p_\theta(\mathbf{z}^t, \mathbf{z}^{t-1}|\mathbf{x}^t, \tilde{\mathbf{x}}^{t-1})$ *as the posterior.* $\log p_\theta(\mathbf{x}^t, \tilde{\mathbf{x}}^{t-1})$ *can be optimized by maximizing a lower bound.*

Proof. In this case, we consider two underlying generative factors (latent variables) $\mathbf{z}^t$, $\mathbf{z}^{t-1}$ for observing the real data $\mathbf{x}^t$ and generated data $\tilde{\mathbf{x}}^{t-1}$, respectively. We then define the latent variable model $p_\theta(\mathbf{x}^t, \tilde{\mathbf{x}}^{t-1}, \mathbf{z}^t, \mathbf{z}^{t-1}) = p_\theta(\mathbf{x}^t, \tilde{\mathbf{x}}^t|\mathbf{z}^t, \mathbf{z}^{t-1})p(\mathbf{z}^t)p(\mathbf{z}^{t-1})$ and its marginal log-likelihood is approximated by a lower bound :

$$\log p_\theta(\mathbf{x}^t, \tilde{\mathbf{x}}^{t-1}) \geq \mathbb{E}_{q_\xi(\mathbf{z}^t|\mathbf{x}^t)}\left[\log \frac{p_\theta(\mathbf{x}^t, \mathbf{z}^t)}{q_\xi(\mathbf{z}^t|\mathbf{x}^t)}\right] + \mathbb{E}_{q_\xi(\mathbf{z}^{t-1}|\tilde{\mathbf{x}}^{t-1})}\left[\log \frac{p_\theta(\tilde{\mathbf{x}}^{t-1}, \mathbf{z}^{t-1})}{q_\xi(\mathbf{z}^{t-1}|\tilde{\mathbf{x}}^{t-1})}\right] \tag{13}$$

We define the above equation as $\mathcal{L}(\theta, \xi; \mathbf{x}^t, \tilde{\mathbf{x}}^{t-1})$.

Lemma 3. *From the Theorems 2 and 3, we can derive a lower bound on the sample log-likelihood at t-th task learning, as expressed by:*

$$\begin{aligned}\log p_\theta(\mathbf{x}^1, .., \mathbf{x}^t) &\geq \mathcal{L}(\theta, \xi; \mathbf{x}^1, .., \mathbf{x}^t) \geq \mathcal{L}(\theta, \xi; \mathbf{x}^t, \tilde{\mathbf{x}}^{t-1}) \\ &- \mathbf{W}(v, v') - \sqrt{2\log\left(\frac{1}{\delta}\right)/a'}\left(\sqrt{\frac{1}{n}} + \sqrt{\frac{1}{n'}}\right) - D^*\end{aligned} \tag{14}$$

where $\nu \in \mathbf{R}^s, \nu' \in \mathbf{R}^s$ *are formed by* n *and* n' *numbers of samples drawn from* $p(\mathbf{x}^t)p(\tilde{\mathbf{x}}^{t-1})$ *and* $\prod_i^t p(\mathbf{x}^i)$*, respectively, where* n *and* n' *denote the sample size.*

Lemma 2 provides an explicit way to investigate how information is lost through GRMs during the lifelong learning process. Meanwhile, Lemma 3 derives the evidence lower bound (ELBO) of the sample log-likelihood. All GRM approaches, based on VAE and GAN architectures, can be explained through this theoretical analysis. However, GAN based approaches lack inference mechanisms, such as the one provided in Lemma 3. This motivates us to develop a new lifelong learning approach utilizing the advantages of both GANs and VAEs, enabling the generation of data and the log-likelihood estimation abilities.

5 The two-step latent variables optimization over time

In the following, inspired by the theoretical analysis from Section 4, we introduce a two-step optimization algorithm which besides training a powerful generative replay network it also induces latent representations. Our algorithm is different from existing hybrid models which would only train generator and inference models within a single optimization function [36] or learn an optimal coupling between generator and inference model by using adversarial learning [11, 15, 16, 36, 37, 42, 43, 61]. The proposed algorithm contains two independent optimization paths, namely "wake" and "dreaming" phases. In the following we explain the application of the proposed Lifelong VAEGAN in supervised, semi-supervised and unsupervised learning.

5.1 Supervised learning

In the "wake" phase, by considering the Definitions 1 and 2, the refined distribution $p(\tilde{\mathbf{x}}^t)$ is trained to approximate $p(\tilde{\mathbf{x}}^{t-1}, \mathbf{x}^t)$ by minimizing the Wasserstein distance :

$$\min_G \max_D \mathcal{L}^G_{GAN}(\theta_t, \omega_t) \triangleq \mathbb{E}_{p(\mathbf{z}),p(\mathbf{c}),p(\mathbf{a})}[D(G(\mathbf{c}, \mathbf{z}, \mathbf{a}))] - \mathbb{E}_{p(\tilde{\mathbf{x}}^{t-1})p(\mathbf{x}^t)}[D(\mathbf{x})], \tag{15}$$

where we omit the penalty term, weighted by λ, for the sake of simplification.

In the "dreaming" phase, we maximize the sample log-likelihood on the joint distribution of the generated data and the empirical data, associated with a given new task, by maximizing the ELBO :

$$\begin{aligned}\mathcal{L}_{VAE}(\theta_t, \varsigma_t, \varepsilon_t, \delta_t) \triangleq \mathbb{E}_{q_{\varsigma,\varepsilon,\delta}(\mathbf{z},\mathbf{a},\mathbf{c}|\mathbf{x}^t)}\left[\log \frac{p_\theta(\mathbf{x}^t|\mathbf{z},\mathbf{a},\mathbf{c})}{q_{\varsigma,\varepsilon,\delta}(\mathbf{z},\mathbf{a},\mathbf{c}|\mathbf{x}^t)}\right] \\ + \mathbb{E}_{q_{\varsigma,\varepsilon,\delta}(\mathbf{z},\mathbf{a},\mathbf{c}|\tilde{\mathbf{x}}^{t-1})}\left[\log \frac{p_\theta(\tilde{\mathbf{x}}^{t-1}|\mathbf{z},\mathbf{a},\mathbf{c})}{q_{\varsigma,\varepsilon,\delta}(\mathbf{z},\mathbf{a},\mathbf{c}|\tilde{\mathbf{x}}^{t-1})}\right].\end{aligned} \tag{16}$$

This loss function is used to train both Generator and Inference models. After training, the Inference model $q_\delta(\mathbf{c}|\mathbf{x})$ can be used for classification.

5.2 Semi-supervised learning

Let us consider that we only have a small subset of labelled data from each database, while the rest of data is unlabelled. For the labelled data, we derive

the objective function without the inference model $q_\delta(\mathbf{c}|\mathbf{x})$ in the "dreaming" phase as :

$$\mathcal{L}^S_{VAE}(\theta_t, \varsigma_t, \varepsilon_t, \delta_t) \triangleq \sum\nolimits^2 \mathbb{E}_{q_\varsigma(\mathbf{z}|\mathbf{x}),q_\varepsilon(\mathbf{a}|\mathbf{x}),p(\mathbf{y})}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{a},\mathbf{y})] - D_{KL}[q_\varsigma(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] - \mathbb{E}_{q_\varsigma(\mathbf{z}|\mathbf{x})} D_{KL}[q_\varepsilon(\mathbf{a}|\mathbf{z})||p(\mathbf{a}|\mathbf{z})] - D_{KL}[q_\delta(\mathbf{c}|\mathbf{x})||p(\mathbf{c})], \tag{17}$$

where $\sum^2$ denotes the estimation of ELBO on the joint model log-likelihood. In addition, we model the unlabeled data samples by using $\mathcal{L}_{VAE}(\theta_t, \varsigma_t, \varepsilon_t, \delta_t)$, where the discrete variable $\mathbf{c}$ is sampled from the Gumbel-softmax distribution whose probability vector is obtained by the encoder $q_\delta(\mathbf{c}|\mathbf{x})$. The full semi-supervised loss used to train the hybrid model is defined as:

$$\mathcal{L}^{Semi}_{VAE} \triangleq \mathcal{L}^S_{VAE} + \beta \mathcal{L}_{VAE}, \tag{18}$$

where β is used to control the importance of the unsupervised learning when compared with the component associated with supervised learning. In addition, the entropy loss $\mathcal{L}_c(\delta)$ is also performed with the labeled data samples in order to enhance the prediction ability of $q_\delta(\mathbf{c}|\mathbf{x})$.

5.3 Unsupervised learning

We also employ the proposed hybrid VAE-GAN model for the lifelong unsupervised learning setting, where we do not have the class labels for each task. Similarly to the supervised learning framework, we minimize the Wasserstein distance between $p(\tilde{\mathbf{x}}^t)$ and $p(\tilde{\mathbf{x}}^{t-1}, \mathbf{x}^t)$:

$$\min_G \max_D \mathcal{L}^U_{GAN}(\theta_t, \omega_t) \triangleq \mathbb{E}_{p(\mathbf{z}),p(\mathbf{a})}[D(G(\mathbf{z},\mathbf{a}))] - \mathbb{E}_{p(\tilde{\mathbf{x}}^{t-1})p(\mathbf{x}^t)}[D(\mathbf{x})]. \tag{19}$$

In the "dreaming" phase, we train the generator and inference models using:

$$\mathcal{L}^U_{VAE}(\theta_t, \varsigma_t, \varepsilon_t) \triangleq \sum\nolimits^2 \mathbb{E}_{q_\varsigma(\mathbf{z}|\mathbf{x}),q_\varepsilon(\mathbf{a}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{a})] - D_{KL}[q_\varsigma(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] - \mathbb{E}_{q_\varsigma(\mathbf{z}|\mathbf{x})} D_{KL}[q_\varepsilon(\mathbf{a}|\mathbf{z})||p(\mathbf{a}|\mathbf{z})], \tag{20}$$

where the Generator is conditioned on only two variables, $\mathbf{z}$ and $\mathbf{a}$. For learning disentangled representations, we employ the Minimum Description Length (MDL) principle [7, 57] and replace the second term from equation (20) by $\gamma|D_{KL}[q_\varsigma(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] - C|$, where γ and C are a multiplicative and a linear constant used for controling the degree of disentanglement.

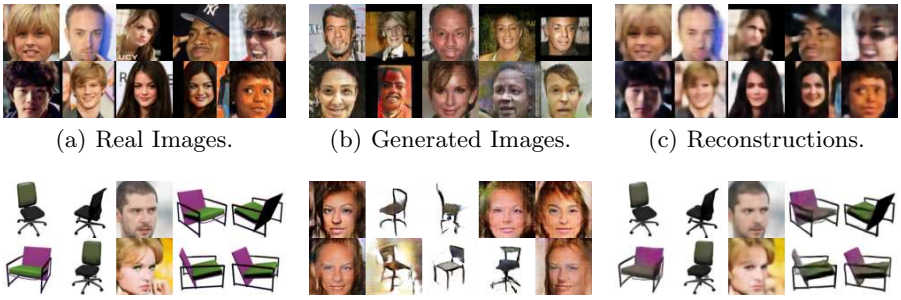
6 Experimental results

6.1 Lifelong unsupervised learning

In this section, we investigate how the proposed Lifelong VAEGAN (L-VAEGAN) model learns meaningful and interpretable representations in images from various databases under the unsupervised lifelong learning setting.

Reconstruction and Interpolation results following lilfeong learning. We train the L-VAEGAN model using the loss functions $\mathcal{L}^U_{GAN}$ and $\mathcal{L}^U_{VAE}$ from

equations (19) and (20), which contain adversarial and VAE learning terms, respectively, and we consider a learning rate of 0.001. The generation and reconstruction results are presented in Figs. 2a-c, for the lifelong learning of CelebA [39] to CACD [9], and in Figs. 2d-f for CelebA to 3DChair [6]. From these results, the inference model works well for both discrete and continuous latent variables.



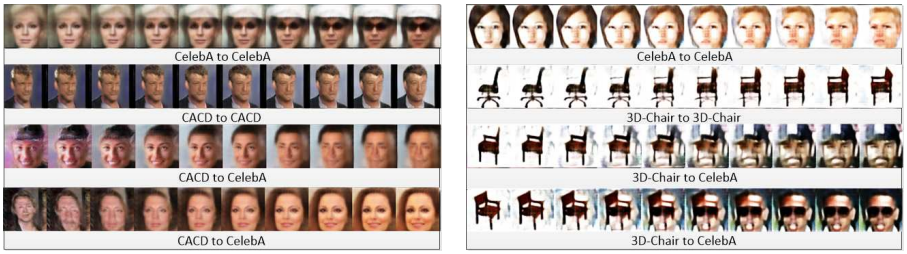
(a) Real Images. (b) Generated Images. (c) Reconstructions.

(d) Real Images. (e) Generated Images. (f) Reconstructions.

Fig. 2. Image reconstructions and generated after the CelebA to CACD (top row) and CelebA to 3DChair (bottom row) lifelong learning.

In the following we perform data interpolation experiments under the lifelong learning setting in order to evaluate the manifold continuity. We call lifelong interpolation when the interpolation is performed between multiple domains, by considering data from different databases, under the lifelong learning setting. We randomly select two images and then infer their discrete $\mathbf{a}$ and continuous $\mathbf{z}$ latent variables by using the inference model. The interpolation results are shown in Figs. 3-a and 3-b, for CelebA to CACD and CelebA to 3D-chair lifelong learning, respectively. The first two rows show the interpolation results in images from the same database, while the last two rows show the interpolations of images from two different databases. From the images from the last two rows of Fig. 3-b we observe that a chair is smoothly transformed into a human face, where its seat and backside are smoothly changed into the eyes and hair of a person. This shows that the L-VAEGAN model can learn the joint latent space of two completely different data configurations.

Lifelong Disentangled Representations. We train the L-VAEGAN model under the CelebA to 3D-Chairs lifelong learning by adapting the loss functions from (19) and (20) in order to achieve unsupervised disentangled representations, as mentioned in Section 5.3. We consider the multiplicative parameter $\gamma = 4$, while increasing the linear one C from 0.5 to 25.0 during the training. After the training, we change one dimension of a continuous latent representation $\mathbf{z}$, inferred by using the inference model for a given input and then recover it back in the visual data space by using the generator. The disentangled results are presented in Fig 4 which indicates changes in the skin, gender, narrowing of the face, chair size, face pose and chairs' style. These results show that the L-VAEGAN hybrid model can discover different disentangled representations in both CelebA and 3D chair databases.



(a) CelebA to CACD learning. (b) CelebA to 3D-chair learning.

Fig. 3. Interpolation results after lifelong learning.

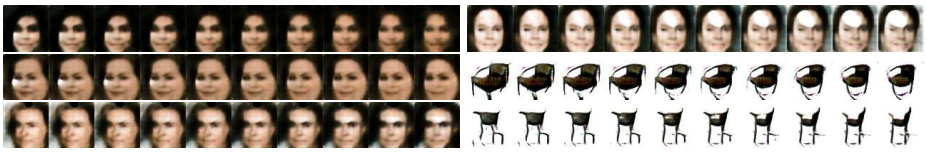
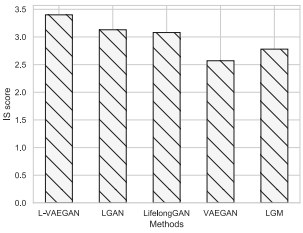


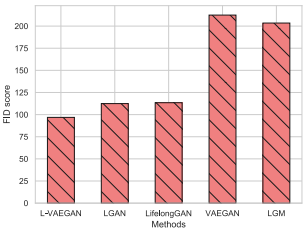
Fig. 4. Manipulating latent variables on CelebA and 3D-chair datasets, under the CelebA to 3D-chairs lifelong learning. We change a single latent variable from -3.0 to 3.0 while fixing all the others. From left to right and top to bottom, we can see changes in skin, gender, narrowing of the face, chair size, face pose and the style of chairs.

Table 1. Quantitative evaluation of the representation learning ability.

The lifelong learning of MNIST and Fashion				
Methods	Lifelong	Dataset	Rec	Acc
L-VAEGAN	M-F	MNIST	**4.75**	**92.53**
LGM [50]	M-F	MNIST	7.18	91.26
VAEGAN [43]	M-F	MNIST	6.54	91.87
L-VAEGAN	M-F	Fashion	**17.44**	**67.66**
LGM [50]	M-F	Fashion	18.33	66.17
VAEGAN [43]	M-F	Fashion	17.03	67.23
L-VAEGAN	F-M	MNIST	**4.92**	**93.29**
LGM [50]	F-M	MNIST	7.18	91.26
VAEGAN [43]	F-M	MNIST	5.52	92.16
L-VAEGAN	F-M	Fashion	**13.16**	**66.97**
LGM [50]	F-M	Fashion	18.83	62.53
VAEGAN [43]	F-M	Fashion	16.57	64.98



(a) IS evaluation.



(b) FID evaluation.

Fig. 5. Quality evaluation.

Quantitative Evaluation. In the following we evaluate numerically the image representation ability for the proposed approach. We consider the reconstruction (Rec) as MSE, and classification accuracy (Acc), as in [67]. The classification accuracy is calculated by a classifier which was trained on the generated images and evaluated on unseen testing samples. We provide the results in Table 1, where L-VAEGAN learns firstly the database MNIST followed by Fashion, denoted as M-F, while F-M represents the learning of the same databases in reversed order.

For comparison we consider LGM [50] and VAEGAN [43], which is one of the best known hybrid models enabled with an inference mechanism. We implement VAEGAN using GRM in order to prevent forgetting. From Table 1 we can see that L-VAEGAN achieves the best results.

We also use the Inception score (IS) [59] and Fréchet Inception Distance (FID) [24] for evaluating the quality of the generated images. We consider using the IS score for evaluating the generated images when training Cifar10 [34] to MNIST lifelong learning and the FID score for CelebA to CACD lifelong learning. Figs. 5-a and 5-b show the IS and FID resuls, respectively, where a lower FID and a higher IS indicate better quality images and where we compare with four lifelong learning approaches : LGAN [60], LifelongGAN [67], VAEGAN [43] and LGM [50]. LifelongGAN [67] requires to load the previously learnt real data samples in order to prevent forgetting them when it is applied in general generation tasks (no conditional images are available). The results indicate that GAN-based lifelong approaches achieve better scores than VAE based methods because VAEs usually generate blurred images. The proposed approach not only produces higher-quality generative replay images but also learns representations of data that other GAN-based lifelong approaches can not. When compared with the VAE-based methods, the proposed approach yields higher-quality reconstructed and generated images.

6.2 Lifelong supervised learning

We compare L-VAEGAN with various other methods under the lifelong supervised setting. LGAN [60] typically trains a classifier (called Solver) on both the images generated by the GAN and the training data samples from the current task. We also consider an auxiliary classifier for LGM [50] (which cannot be used directly as a classifier) by training it on the mixed dataset consisting of images generated by the LGM and the training samples of the current task.

Table 2. The classification results for the lifelong learning of MNIST and Fashion.

Dataset	Lifelong	L-VAEGAN	LGAN[60]	LGM[50]	EWC [33]	Transfer	MeRGANs [64]
MNIST	M-F	**98.76**	98.41	97.29	37.7	40.63	98.34
MNIST	F-M	98.77	98.32	98.85	**99.12**	98.25	98.27
Fashion	M-F	**92.01**	91.42	91.71	91.38	91.01	91.12
Fashion	F-M	**89.24**	89.15	86.05	54.53	37.92	88.86

We train the L-VAEGAN model under the MNIST to SVHN [45] and MNIST to Fashion lifelong learning tasks, respectively. The classification accuracy of various methods is reported in Tables 2 and 3. We observe that the replay generative images can prevent forgetting and the quality of performing on the previous tasks is depending on the capability of the Generator. For instance, the classification accuracy of previous tasks for LGM is a bit lower than for the other

Table 3. The classification results for the lifelong learning of MNIST and SVHN.

Dataset	Lifelong	L-VAEGAN	LGAN[60]	LGM[50]	EWC [33]	Transfer	MeRGANs [64]
MNIST	M-S	**96.63**	96.59	96.59	32.63	4.70	96.46
MNIST	S-M	98.55	98.42	98.30	**99.03**	98.45	98.38
SVHN	M-S	81.97	80.77	80.10	89.71	**89.90**	79.67
SVHN	S-M	**80.99**	76.76	80.97	37.69	3.60	80.85

GAN-based methods. The reason for this is that LGM uses VAEs as a generator and therefore can not generate high-quality images when compared to GANs. The classification results demonstrate that the proposed approach can be used in the lifelong supervised learning with a good performance.

Table 4. Semi-supervised classification error results on MNIST database, under the MNIST to Fashion lifelong learning.

Methods	Lifelong?	Error
L-VAEGAN	Yes	4.34
LGAN [60]	Yes	5.46
Neural networks (NN) [31]	No	10.7
(CNN) [31]	No	6.45
TSVM [31]	No	5.38
CAE [31]	No	4.77
M1+TSVM [31]	No	4.24
M2 [31]	No	3.60
M1+M2 [31]	No	2.40
Semi-VAE [44]	No	2.88

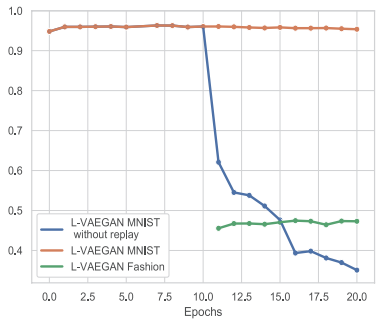


Fig. 6. The accuracy during the semi-supervised training on the testing data samples during the lifelong learning. The model is trained for 10 epochs for each task.

6.3 Lifelong semi-supervised learning

The semi-supervised learning framework for L-VAEGAN is presented in Section 5.2. During the semi-supervised training we consider only a small number of labelled images from each database. In the following experiments we divide MNIST and Fashion datasets into two subsets each, representing labelled and unlabeled data. We consider a total of 1,000 and 10,000 labelled images for MNIST and Fashion datasets, respectively, with an equal number of data in each class for the labelled set. We train the proposed L-VAEGAN model with 10 epochs for learning each database and the resulting classification plots are shown in Fig. 6. We observe that without generative replay samples, the model under the semi-supervised setting suffers from catastrophic forgetting.

The classification results for lifelong learning when using L-VAEGAN compared to other semi-supervised learning methods are provided in Table 4. These results show that the proposed approach outperforms LGAN [60], under the semi-supervised learning setting, and achieves competitive results when compared to the state-of-the art models which are not trained using lifelong learning.

6.4 Analysis

The plot of the classification accuracy for all testing samples during the MNIST to Fashion lifelong learning is provided in Fig. 7-a, where the first 10 training epochs correspond to learning the MNIST task, while the next 10 epochs are used for learning the Fashion database. We observe that the performance of previous tasks is maintained when considering the GRM. However, without replaying data, the model learned from the previous tasks quickly forgets the past knowledge while learning new tasks, as indicated by a significant performance drop, as observed in Fig. 7-a. Besides, unlike LGAN [60], the proposed L-VAEGAN is able to provide inference mechanisms, benefiting many downstream tasks, as shown in the results from Fig. 2.

In the following, in Fig. 7-b we provide numerical results for the generalization bounds for the GRM, described in Section 4, where risk1 and risk2 denote $E(h^1(\mu_1))$ and $E(h^1(\mu_{1'}))$, respectively. We find that $E(h^1(\mu_1))$ is very close to $E(h^1(\mu_{1'}))$ and still a bound on $E(h^1(\mu_{1'}))$ during the training. Fig. 7-c provides the numerical results for the observed risks under the MNIST-MNIST lifelong learning. We observe that $E(h^1(\mu_1))$ is still a bound on $E(h^1(\mu_{1'}))$ and this bound is gradually increased during the training. The reason for this is that the model is gradually adapting $p(\tilde{\mathbf{x}}^1)$ and the bound depends on the distance between $p(\tilde{\mathbf{x}}^1)$ and $p(\mathbf{x}^1)$.

(a) Forgetting curves (b) MNIST-Fashion (c) MNIST-MNIST

Fig. 7. Classification accuracy and observed risks during the lifelong learning.

7 Conclusion

A novel hybrid model for lifelong representation learning, called Lifelong VAE-GAN (L-VAEGAN) is proposed in this research study. A two-step optimization algorithm, which can also induce higher-quality generative replay samples and learn informative latent representations over time, is used to train L-VAEGAN. The results indicate that L-VAEGAN model is able to discover disentangled representations from multiple domains under the lifelong learning setting. More importantly, L-VAEGAN automatically learns semantic meaningful shared latent variables across different domains, which allows to perform cross-domain interpolation and inference.

References

1. Achille, A., Eccles, T., Matthey, L., Burgess, C., Watters, N., Lerchner, A., Higgins, I.: Life-long disentangled representation learning with cross-domain latent homologies. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 9873–9883 (2018)
2. Aljundi, R., Caccia, L., Belilovsky, E., Caccia, M., Lin, M., Charlin, L., Tuytelaars, T.: Online continual learning with maximal interfered retrieval. In: Proc. Neural Inf. Proc. Systems (NIPS), arXiv preprint arXiv:1908.04742 (2019)
3. Aljundi, R., Chakravarty, P., Tuytelaars, T.: Expert gate: Lifelong learning with a network of experts. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 3366–3375 (2017)
4. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. In: Proc. Neural Inf. Proc. Systems (NIPS), arXiv preprint arXiv:1903.08671 (2019)
5. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proc. Int. Conf. on Machine Learning (ICML). pp. 214–223 (2017)
6. Aubry, M., Maturana, D., Efros, A.A., Russell, B.C., Sivic, J.: 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). pp. 3762–3769 (2014)
7. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in β-vae. In: Proc. NIPS Workshop on Learning Disentangled Represen., arXiv preprint arXiv:1804.03599 (2017)
8. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P., Torr, P.H.S., Ranzato, M.: On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486 (2019)
9. Chen, B.C., Chen, C.S., Hsu, W.H.: Cross-age reference coding for age-invariant face recognition and retrieval. In: Proc. European Conf on Computer Vision (ECCV), vol. LNCS 8694. pp. 768–783 (2014)
10. Chen, L., Dai, S., Pu, Y., Li, C., Su, Q., Carin, L.: Symmetric variational autoencoder and connections to adversarial learning. In: Proc. Int. Conf. on Artificial Intel. and Statistics (AISTATS) 2018, vol. PMLR 84. pp. 661–669 (2018)
11. Chen, L., Dai, S., Pu, Y., Li, C., Su, Q., Carin, L.: Symmetric variational autoencoder and connections to adversarial learning. arXiv preprint arXiv:1709.01846 (2017)
12. Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 2615–2625 (2018)
13. Chen, Z., Ma, N., Liu, B.: Lifelong learning for sentiment classification. In: Proc. of the Annual Meeting of the Assoc. for Comp. Linguistics and Int. Joint Conf. on Natural Language Processing. pp. 750–756 (2015)
14. Dai, W., Yang, Q., Xue, G.R., Yu, Y.: Boosting for transfer learning. In: Proc. Int Conf. on Machine Learning (ICML). pp. 193–200 (2007)
15. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. In: Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1605.09782 (2017)
16. Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., Courville, A.: Adversarially learned inference. In: Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1606.00704 (2017)

17. Dupont, E.: Learning disentangled joint continuous and discrete representations. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 710–720 (2018)
18. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2147–2154 (2014)
19. Fagot, J., Cook, R.G.: Evidence for large long-term memory capacities in baboons and pigeons and its implications for learning and the evolution of cognition. Proc. of the National Academy of Sciences (PNAS) **103**(46), 17564–17567 (2006)
20. Gao, S., Brekelmans, R., ver Steeg, G., Galstyan, A.: Auto-encoding total correlation explanation. In: Proc. Int. Conf. on Artificial Intel. and Statistics (AISTATS) 2018, vol. PMLR 89. pp. 1157–1166 (2019)
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 2672–2680 (2014)
22. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 5767–5777 (2017)
23. Gumbel, E.J.: Statistical theory of extreme values and some practical applications:a series of lectures (1954)
24. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proc. Advances in Neural Information Processing Systems (NIPS). pp. 6626–6637 (2017)
25. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: β-VAE: Learning basic visual concepts with a constrained variational framework. In: Proc. Int. Conf. on Learning Representations (ICLR) (2017)
26. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: Proc. NIPS Deep Learning Workshop, arXiv preprint arXiv:1503.02531 (2014)
27. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with Gumbel-Softmax. In: Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1611.01144 (2017)
28. Jeong, Y., Song, H.O.: Learning discrete and continuous factors of data via alternating disentanglement. In: Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 97. pp. 3091–3099 (2019)
29. Jung, H., Ju, J., Jung, M., Kim, J.: Less-forgetting learning in deep neural networks. arXiv preprint arXiv:1607.00122 (2016)
30. Kim, H., Mnih, A.: Disentangling by factorising. In: Proc. of Int. Conf. on Machine Learning (ICML), vol. PMLR 80. pp. 2649–2658 (2018)
31. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 3581–3589 (2014)
32. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114 (2013)
33. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proc. of the National Academy of Sciences (PNAS) **114**(13), 3521–3526 (2017)
34. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. rep. (2009)

35. Kumar, A., Sattigeri, P., Balakrishnan, A.: Variational inference of disentangled latent concepts from unlabeled observations. In: Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1711.00848 (2018)
36. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: Proc. Int Conf. on Machine Learning (ICML). pp. 1558–1566 (2015)
37. Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R., Carin, L.: Alice: Towards understanding adversarial learning for joint distribution matching. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 5495–5503 (2017)
38. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans. on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2017)
39. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proc. of IEEE Int. Conf. on Computer Vision (ICCV). pp. 3730–3738 (2015)
40. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. In: Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1611.00712 (2016)
41. Maddison, C.J., Tarlow, D., Minka, T.: A* sampling. Proc. Advances in Neural Inf. Proc. Systems (NIPS) p. 1–10 (2014)
42. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. In: Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1511.05644 (2016)
43. Mescheder, L., Nowozin, S., Geiger, A.: Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In: Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 70. pp. 2391–2400 (2017), https://arxiv.org/abs/1701.04722
44. Narayanaswamy, S., Paige, T.B., Van de Meent, J.W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., Torr, P.: Learning disentangled representations with semi-supervised deep generative models. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 5925–5935 (2017)
45. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
46. Nowozin, S., Cseke, B., Tomioka, R.: f-gan: Training generative neural samplers using variational divergence minimization. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 271–279 (2016)
47. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113**, 54–71 (2019)
48. Polikar, R., Upda, L., Upda, S.S., Honavar, V.: Learn++: An incremental learning algorithm for supervised neural networks. IEEE Trans. on Systems Man and Cybernetics, Part C **31**(4), 497–508 (2001)
49. Pu, Y., Wang, W., Henao, R., L., C., Gan, Z., Li, C., Carin, L.: Adversarial symmetric variational autoencoder. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 4333–4342 (2017)
50. Ramapuram, J., Gregorova, M., Kalousis, A.: Lifelong generative modeling. In: Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1705.09847 (2017)
51. Rannen, A., Aljundi, R., Blaschko, M., Tuytelaars, T.: Encoder based lifelong learning. In: Proc. of the IEEE Int. Conf. on Computer Vision (ICCV). pp. 1320–1328 (2017)

52. Rao, D., Visin, F., Rusu, A.A., Teh, Y.W., Pascanu, R., Hadsell, R.: Continual unsupervised representation learning. In: Proc. Neural Inf. Proc. Systems (NIPS), arXiv preprint arXiv:1910.14481 (2019)
53. Redko, I., Habrard, A., Sebban, M.: Theoretical analysis of domain adaptation with optimal transport. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 737–753. Springer (2017)
54. Ren, B., Wang, H., Li, J., Gao, H.: Life-long learning based on dynamic combination model. Applied Soft Computing **56**, 398–404 (2017)
55. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 91–99 (2015)
56. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 32. pp. 1278–1286 (2014)
57. Rissanen, J.: Modeling by shortest data description. Automatica **14**(5), 465–471 (1978)
58. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016)
59. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 2234–2242 (2016)
60. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 2990–2999 (2017)
61. Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U., Sutton, C.: Veegan: Reducing mode collapse in gans using implicit variational learning. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 3308–3318 (2017)
62. Tessler, C., Givony, S., Zahavy, T., Mankowitz, D.J., Mannor, S.: A deep hierarchical approach to lifelong learning in Minecraft. In: Proc. AAAI Conf. on Artificial Intelligence. pp. 1553–1561 (2017)
63. Wang, X., Zhang, R., Sun, Y., Qi, J.: KDGAN: knowledge distillation with generative adversarial networks. In: Proc. Advances in Neural Inf. Proc. Systems (NIPS). pp. 775–786 (2018)
64. Wu, C., Herranz, L., Liu, X., van de Weijer, J., Raducanu, B.: Memory replay GANs: Learning to generate new categories without forgetting. In: Proc. Advances In Neural Inf. Proc. Systems (NIPS). pp. 5962–5972 (2018)
65. Yoon, J., Yang, E., Lee, J., Hwang, S.J.: Lifelong learning with dynamically expandable networks. In: Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1708.01547 (2017)
66. Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M., Mori, G.: Lifelong GAN: Continual learning for conditional image generation. arXiv preprint arXiv:1907.10107 (2019)
67. Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M., Mori, G.: Lifelong gan: Continual learning for conditional image generation. In: Proc. of the IEEE Int. Conf. on Computer Vision (ICCV). pp. 2759–2768 (2019)