LIFELONG TWIN GENERATIVE ADVERSARIAL NETWORKS

Fei Ye and Adrian G. Bors

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

ABSTRACT

In this paper, we propose a new continuously learning generative model, called the Lifelong Twin Generative Adversarial Networks (LT-GANs). LT-GANs learns a sequence of tasks from several databases and its architecture consists of three components: two identical generators, namely the Teacher and Assistant, and one Discriminator. In order to allow for the LT-GANs to learn new concepts without forgetting, we introduce a new lifelong training approach, namely Lifelong Adversarial Knowledge Distillation (LAKD), which encourages the Teacher and Assistant to alternately teach each other, while learning a new database. This training approach favours transferring knowledge from a more knowledgeable player to another player which knows less information about a previously given task.

***Index Terms*—** Lifelong learning, Generative Adversarial Networks (GAN), Teacher-Student learning models.

1. INTRODUCTION

An essential characteristic of human beings intelligence is that of being able to continually learn and acquire new skills and concepts from the world throughout their lifespan [1]. Neural networks trained on a sequence of tasks tend to focus on the latest learnt task and would perform poorly on any other learnt before. This phenomenon is called catastrophic forgetting [2], and it is caused by the fact that network's parameters are overwritten each time when training on a new task. A solution proposed to relieve catastrophic forgetting was to impose constraints while learning the tasks associated with a new database [3]. This is achieved by including a regularization term in the objective function, penalizing the change in the network weights when learning a new task. Other solutions focused on dynamically increasing the number of neurons and network layers in order to be able to store novel information, [4]. Most of these approaches require task labels, or the knowledge of the task boundaries, which is not always feasible. Moreover, these approaches are focused on the supervised learning setting while in this paper we address the more challenging problem of unsupervised learning, [5].

Generative Adversarial Nets (GANs) [6] can relieve catastrophic forgetting by being trained in a self-supervised fashion. Retraining with generative replay is achieved by building new training sets comprising of data generated by a GAN, which are added to a given real dataset from the current task [7], or by preserving and freezing the model's parameters after each dataset switch. However, applying these approaches in practical applications, that would favour a memory-efficient and end-to-end learning manner, is challenging. Moreover, existing GAN based lifelong learning models lack inference mechanisms and therefore can not capture complex structures behind data.

The proposed lifelong learning model, is named Lifelong Learning GANs (LT-GANs). LT-GANs relieves catastrophic forgetting by accumulating knowledge through a twin Teacher-Assistant network in the context of adversarial learning. This learning process favours transferring knowledge from a more knowledgeable player to its twin player during the lifelong learning, while the memory size is rather small and kept fixed. We further implement a Teacher-Student framework considering the LT-GANs as a Teacher network in order to learn data representations over time.

This research study has the following contributions:

- A new lifelong learning model, LT-GANs, which aims to learn a sequence of tasks from a set of databases.
- We introduce the Lifelong Adversarial Knowledge Distillation (LAKD), an end-to-end training algorithm for accumulating knowledge across tasks.
- The LT-GANs model is extended to a Teacher-Student framework for learning data representations over time.

2. RELATED WORKS

There are two categories of lifelong learning approaches : memory-based systems [8], and generative modelling [5]. The former category of models uses a small buffer to store some data samples for each task which are then used to minimize the negative backward transfer when updating the network's parameters when learning a new task [9]. However, such approaches require a significant computation processing [8], when the number of tasks to be learnt increases. The generative modelling approaches would usually use a generator such as a Generative Adversarial Network (GAN) [6] or a Variational Autoencoder (VAE) [10] to reproduce previously learnt data samples before learning the next task. These generated data samples are then mixed with samples drawn from the current database, to form a new training data set for the model. Moreover, existing GAN based lifelong approaches [11] can not learn inference models, which prevent

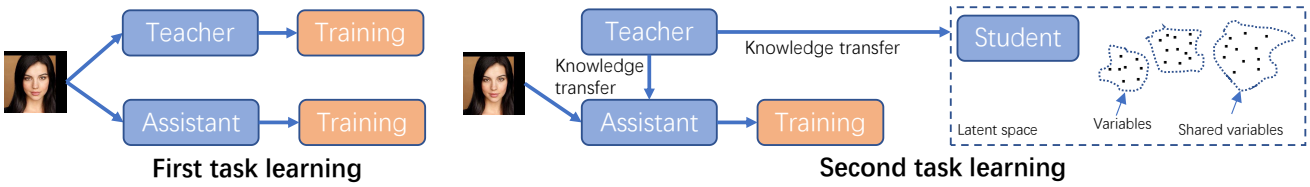


Fig. 1. Diagram showing the learning flow for LT-GANs. The Teacher and Assistant are trained jointly during the first task, while they exchange their roles, teaching each other in turns, when learning the following tasks.

their usefulness for a wide range of applications. In contrast, VAE based lifelong approaches [12] are able to capture cross-domain representations over several tasks but lead to poor performance when learning databases of high complexity, given that VAEs used as generative replay networks tend to produce rather blurred images.

Another category of related works is based on the coupled [13] or dual generative models [14]. CoGANs [13] consists of a pair of GANs which share their parameters for generator and discriminator networks. DualGAN [14] has a similar network architecture with CoGAN, with the difference that DualGAN aims to learn an image-to-image translation framework while CoGAN aims to learn a joint distribution without accessing a tuple of corresponding images. The Twin Auxiliary Classifiers GANs (TAC-GAN) [15] introduces a classifier as a new player to interact with the GAN's generator and discriminator in order to enforce the diversity in the generated data. In this paper we propose the Lifelong Adversarial Knowledge Distillation (LAKD) as the training algorithm which enables the LT-GANs model to learn new tasks without forgetting.

3. LIFELONG TWIN GENERATIVE ADVERSARIAL NETWORKS (LT-GANS)

Most lifelong learning models are applied in the context of supervised learning. The proposed Lifelong Twin Generative Adversarial Network (LT-GANs) consists of three components: two identical generators and a discriminator network. One of the generators is named the Teacher while the other one is the Assistant. LT-GANs training procedure is shown in the diagram from Fig. 1. Let $\mathbf{z}$ represent a random noise vector sampled from a Normal distribution $\mathcal{N}(0, \mathbf{I})$. The two generators, Teacher and Assistant, are parameterized by two identical neural networks $G_{\theta_T}(\mathbf{z})$ and $G_{\theta_A}(\mathbf{z})$, which aim to generates images $\mathbf{x}'_t$ and $\mathbf{x}'_a$ by taking the random noise vector $\mathbf{z}$ as input. The learning goal of LT-GANs in the first task is similar to that in GAN [6], by minimizing the distance between the data distribution and the distribution of the generated data. In this paper, we consider minimizing the Wasserstein distance, [16] :

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

where Θ represents a set of 1-Lipschitz functions. $p(\mathbf{x}_T)$ and $p(\mathbf{x}_A)$ represent the generator distributions for the Teacher and Assistant networks, $G_{\theta_T}(\mathbf{z})$ and $G_{\theta_A}(\mathbf{z})$, respectively, while $D(\cdot)$ represents the discriminator network. We introduce a gradient penalty term (momentum) [16], defined by λ, in order to enforce the Lipschitz constraint, resulting in:

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

$\mathbb{P}_{\tilde{\mathbf{x}}_T}$ and $\mathbb{P}_{\tilde{\mathbf{x}}_A}$ are defined by sampling uniformly along straight lines between pairs of data from the given distribution $p(\mathbf{x}^1)$, and the distribution generated by the Teacher net, and those sampled from $p(\mathbf{x}^1)$ and the Assistant's distribution.

Traditional knowledge distribution approaches normally train a classifier on the predictions made by another classifier [17]. Some recent studies have proposed to learn a single model from an ensemble of networks in order to achieve a higher performance while requiring a lighter computational cost. These approaches, however, would require real data samples as well as supervision signals drawn from a single domain for knowledge distillation, which is a serious challenge for lifelong learning. In this paper, we propose the Lifelong Adversarial Knowledge Distillation (LAKD) for training LT-GANs. The main idea of LAKD is to encourage one of the generators to be a Teacher and to transfer its knowledge to another generator, which is the Assistant, during the lifelong learning. Let us assume that the Teacher and Assistant have been trained on the first task. When learning the second task, the Teacher has its parameters fixed and is seen as the knowledgeable source, while the Assistant learns data samples drawn from the Teacher's distribution as well as the given real data. During the third task learning, the Teacher and Assistant exchange their roles and the Assistant becomes now the Teacher and has its weights fixed, while transferring knowledge to the former Teacher which becomes now the Assistant. The LAKD loss 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 $\mathbf{x}$ are uniformly sampled from $p(\mathbf{x}^k)$ and $p(\mathbf{x}_T^{k-1})$, where $p(\mathbf{x}_T^{k-1})$ denotes the distribution $G_{\theta_T^{k-1}}(\mathbf{z})$, generated by the Teacher, which had been trained on the $(k-1)$-th task. The fake data sample $\mathbf{x}'$ is drawn from $p(\mathbf{x}_A^k)$ which represents the distribution $G_{\theta_A^k}(\mathbf{z})$ generated by the Assistant following its training on the k-th task. Training LT-GANs using

the proposed LAKD has many advantages when compared to other lifelong learning approaches [5, 9]. Firstly, LAKD does not require to load previously learnt data samples [7] while its memory size does not change as the number of tasks increases. Secondly, it does not require to preserve the model's parameters or even a snapshot of these after each task switch, as we have in other generative replay methods [12].

4. THE LIFELONG TEACHER-STUDENT

We consider a Teacher-Student architecture, where the Teacher is represented by the LT-GANs, while the Student is implemented 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 given that it requires integration over the entire latent variable space $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$. Instead, we maximize the evidence lower bound (ELBO) on the sample log-likelihood, as in the Variational Autoencoder (VAE) inference, [10] :

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

where $p_\omega(\mathbf{x}|\mathbf{z})$ is the probability implemented by the decoder and $q_\varepsilon(\mathbf{z}|\mathbf{x})$ is that of the inference model, implemented by a neural network which has Gaussian-specific prior parameters $\{\mu, \sigma\}$ for its last layer's outputs, while D_{KL} is the Kullback-Leibler (KL) divergence. The latent vector $\mathbf{z}$ is sampled using the reparametrisation trick $\mathbf{z} = \mu + \gamma \otimes \sigma$, where γ is random noise drawn from $\mathcal{N}(0, \mathbf{I})$.

For learning cross-domain representations under the lifelong learning, we consider transferring the knowledge from the most knowledgeable generator to the Student network:

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

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 network is more knowledgeable when learning the k-th task. $\mathbf{x}'$ is sampled from $p(\mathbf{x}_Q^{k-1})$. The Student network training is synchronised with that of LT-GANs.

We enable the Student for learning dissentangled representations [18, 19] by penalizing the Kullback-Leibler divergence between the posterior and prior distributions, [20] :

$$\begin{aligned}\log[p(\mathbf{x}^k)p(\mathbf{x}_Q^{k-1})] \geq\ & \mathbb{E}_{\mathbf{z}\sim q_\varepsilon(\mathbf{z}|\mathbf{x}^k)}[\log p_\omega(\mathbf{x}|\mathbf{z})] \\ & - \beta D_{KL}[q_\varepsilon(\mathbf{z}|\mathbf{x}^k)||p(\mathbf{z})] \\ & + \mathbb{E}_{\mathbf{z}\sim q_\varepsilon(\mathbf{z}|\mathbf{x}')}[\log p_\omega(\mathbf{x}|\mathbf{z})] \\ & - \beta D_{KL}[q_\varepsilon(\mathbf{z}|\mathbf{x}')||p(\mathbf{z})] = \mathcal{L}_{Dis}(\omega, \varepsilon)\end{aligned} \tag{6}$$

where $\beta = 1$ corresponds to (5). A large β encourages the independence between latent variables but sacrifices the reconstruction quality. During the experiments we set $\beta = 4$.

5. EXPERIMENTS

We train the proposed lifelong Teacher-Student model, using the ELBO criterion, (5). The results generated after the lifelong learning one database under the CelebA [21] to CACD [22], and CelebA to 3D-Chair database lifelong learning, shown in Fig. 2A and 2B, respectively. Then we interpolate between two latent vectors encoding two different images, from the same database and also from different databases, with the results shown in Fig. 2C and 2D, for the same databases as above. We can observe that an image can be smoothly transformed into another, even when the two images come from two different domains, as when interpolating between a 3D chair and a face image, as shown in the second example from Fig. 2D. These results indicate that the Student module can capture shared and domain-specific generative factors over time.

Dataset	LT-GANs	CURL [5]	LGM [23]
MNIST	**878.92**	887.40	900.18
SVHN	**219.15**	261.08	262.20
Fashion	**365.62**	639.19	642.20
Omniglot	**514.87**	695.68	700.84
Average	**494.64**	620.83	626.35

Table 1. The negative log-likelihood estimation for the lifelong learning of MNIST, SVHN, Fashion and Omniglot.

Tasks	LT-GANs	CURL [5]
First task	**62.85**	155.59
Second task	**59.27**	166.47
Third task	**60.35**	169.28

Table 2. FID score after the lifelong learning of CIFAR10, CIFAR100, Sub1- and Sub2-ImageNet databases.

We evaluate the performance of various lifelong learning models when training LT-GANs using MNIST, SVHN, Fashion and Omniglot databases (MSFO sequence). The negative log-likelihood (NLOG) results, estimated as the reconstruction error plus the KL term, are provided in Table 1. LT-GANs performs better in all tasks, while VAE based lifelong methods, such as CURL [5] and LGM [23] tend to forget previously learnt tasks. We train various models under the lifelong learning of CIFAR10 and other two distinct datasets, called Sub1 and Sub2, which are subsets of the ImageNet database [24]. In Table 2 we evaluate the quality of the generated images by calculating the Fréchet Inception Distance (FID) [25], after each task switch during the lifelong learning.

We also train the proposed Teacher-Student framework under the CelebA to 3D-Chairs by using the loss defined by (6). We then manipulate the latent space by changing a single latent variable while fixing the others. The generated image results when using the LT-GANs model are shown in Figs. 3-a and 3-b, when changing the facial expression of a woman, and the shape of a chair from CelebA and 3D chair databases.

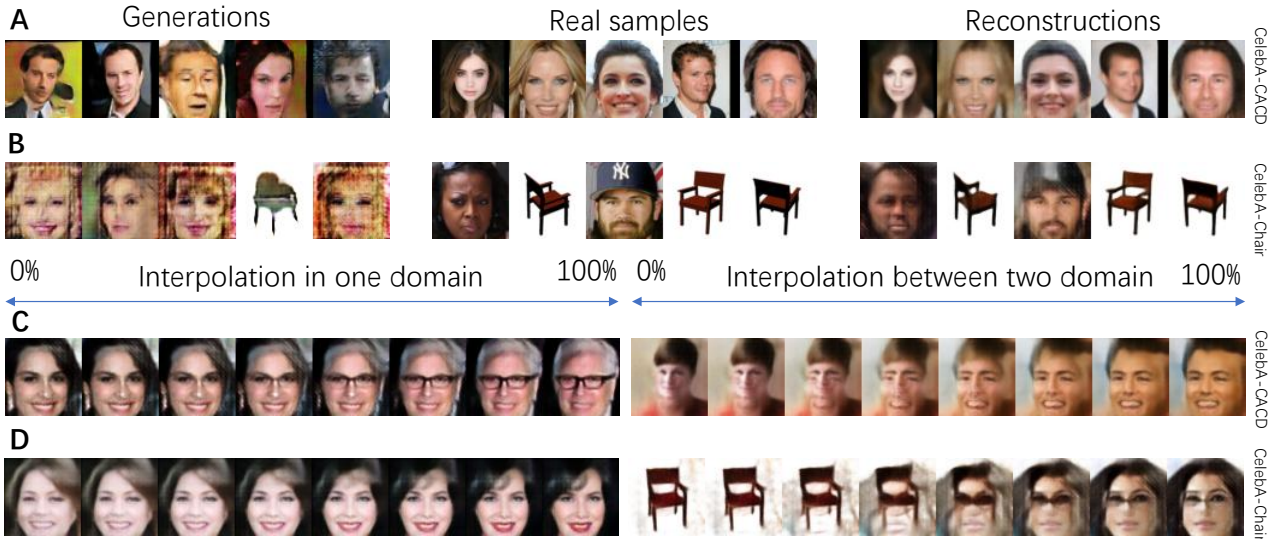
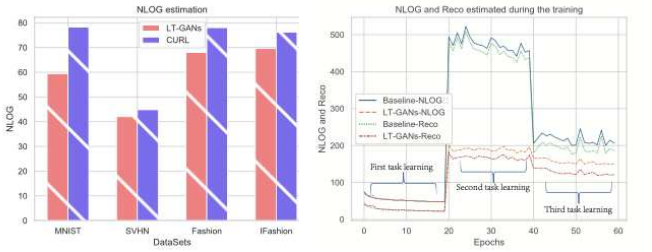


Fig. 2. Image generation results when employing LT-GANs for the lifelong learning of CelebA-to-CACD and CelebA-to-3D Chairs databases are provided in A) and B). Interpolations in the latent space for the same databases are provided in C) and D).

(a) Changing facial expression.

(b) Changing the shape of a chair.

Fig. 3. Disentangled representation results after the lifelong learning of CelebA to 3D-Chairs databases.



(a) NLOG on various datasets. (b) NLOG and MSE on MNIST.

Fig. 4. Analysis results for LT-GANs.

In order to test the robustness of the lifelong generative models we consider fuzzy task boundaries. In these experiments we exchange certain images, from one of the classes, between two databases and create new databases, MNIST-c, SVHN-c and Fashion-c, while preserving the images for the other nine classes. The NLOG results are provided in Table 3, where the proposed Teacher-Student framework still achieves the best results in terms of NLOG image reconstruction.

We also split the database into data sets containing images of a certain class, corresponding to 10 tasks in total. We train various models on one database with 10 tasks and evaluate

Dataset	LT-GANs	CURL [5]	LGM [23]
MNIST-c	**139.98**	437.45	1365.41
SVHN-c	**141.57**	209.70	206.44
Fashion-c	**51.68**	54.72	145.78
Average	**111.08**	233.96	556.13

Table 3. NLOG for the lifelong learning with poorly defined task boundaries on MNIST-c, SVHN-c and Fashion-c.

NLOG on all testing samples, and the results are provided in Fig. 4-a. From this bar-plot it can be observed that the proposed framework outperforms CURL [5]. We investigate the effect of the proposed LAKD loss function by changing β in (6) for the lifelong learning in terms of the log-likelihood and reconstruction errors measured as the MSE error. LAKD loss function plays an important role in overcoming forgetting, according to Fig. 4-b, where the baseline is considered when the Teacher and Assistant are trained jointly in each task.

6. CONCLUSION

We introduce a new lifelong learning LT-GANs model, made up of a dual-generator network, which is trained in a memory-efficient and end-to-end learning manner using the proposed Lifelong Adversarial Knowledge Distillation (LAKD) loss function. We further extend the LT-GANs model into a Teacher-Student framework in order to capture data representations, where the two generators teach alternatively one another, as well as to a Student network. The proposed framework is enabled with the ability to model disentangled representations under the unsupervised lifelong learning setting. It is also shown to generate smooth interpolations between images associated with different databases. The results demonstrate that the proposed framework achieves the state of the art in lifelong unsupervised representation learning.

7. REFERENCES

[1] J. Cichon and W.-B. Gan, "Branch-specific dendritic ca 2+ spikes cause persistent synaptic plasticity," *Nature*, vol. 520, no. 7546, pp. 180–185, 2015.

[2] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[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 Nat. Acad. of Sciences (PNAS)*, vol. 114, no. 13, pp. 3521–3526, 2017.

[4] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," in *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1708.01547*, 2017.

[5] D. Rao, F. Visin, A. A. Rusu, Y. W. Teh, R. Pascanu, and R. Hadsell, "Continual unsupervised representation learning," in *Advances in Neural Inf. Proc. Systems (NIPS), arXiv preprint arXiv:1910.14481*, 2019.

[6] 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. Syst. (NIPS)*, 2014, pp. 2672–2680.

[7] 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.

[8] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," in *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1812.00420*, 2019.

[9] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. Dokania, P. H. S. Torr, and M.'A. Ranzato, "On tiny episodic memories in continual learning," *arXiv preprint arXiv:1902.10486*, 2019.

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

[11] Fei Ye and Adrian G. Bors, "Lifelong learning of interpretable image representations," in *Proc. Int. Conf. on Image Processing Theory, Tools and Applications (IPTA)*, 2020, pp. 1–6.

[12] Fei Ye and Adrian G Bors, "Learning latent representations across multiple data domains using lifelong vaegan," in *Proc. of European Conf. on Computer Vision (ECCV), vol. LNCS 12365*, 2020, pp. 777–795.

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

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

[15] M. Gong, Y. Xu, C. Li, K. Zhang, and K. Batmanghelich, "Twin auxilary classifiers GAN," in *Advances in Neural Inf. Proc. Systems (NIPS)*, pp. 1330–1339. 2019.

[16] 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. 5767–5777.

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

[18] Fei Ye and Adrian G. Bors, "Deep mixture generative autoencoders," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[19] Fei Ye and Adrian G Bors, "Learning joint latent representations based on information maximization," *Information Sciences*, vol. 567, no. 8, pp. 216–236, 2021.

[20] 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 *Int. Conf. on Learning Representations (ICLR)*, 2017.

[21] 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.

[22] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. of European Conf. on Computer Vision (ECCV), vol. LNCS 8694*, 2014, pp. 768–783.

[23] J. Ramapuram, M. Gregorova, and A. Kalousis, "Lifelong generative modeling," in *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1705.09847*, 2017.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[25] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 6626–6637.