Proceedings Paper:
Ye, Fei and Bors, Adrian Gheorghe orcid.org/0000-0001-7838-0021 (2023) Compressing Cross-Domain Representation via Lifelong Knowledge Distillation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 04-10 Jun 2023 IEEE , GRC

https://doi.org/10.1109/ICASSP49357.2023.10096910

COMPRESSING CROSS-DOMAIN REPRESENTATION VIA LIFELONG KNOWLEDGE DISTILLATION

Fei Ye and Adrian G. Bors[*]

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

ABSTRACT

Most Knowledge Distillation (KD) approaches focus on the discriminative information transfer and assume that the data is provided in batches during training stages. In this paper, we address a more challenging scenario in which different tasks are presented sequentially, at different times, and the learning goal is to transfer the generative factors of visual concepts learned by a Teacher module to a compact latent space represented by a Student module. In order to achieve this, we develop a new Lifelong Knowledge Distillation (LKD) framework where we train an infinite mixture model as the Teacher which automatically increases its capacity to deal with a growing number of tasks. In order to ensure a compact architecture and to avoid forgetting, we propose to measure the relevance of the knowledge from a new task for a set of experts making up the Teacher module, guiding each expert to capture the probabilistic characteristics of several similar domains. The network architecture is expanded only when learning an entirely different task. The Student is implemented as a lightweight probabilistic generative model. The experiments show that LKD can train a compressed Student module that achieves the state of the art results with fewer parameters.

Index Terms— Lifelong learning, Generative models, Teacher-Student architecures, Mixtures of experts.

1. INTRODUCTION

Lifelong learning (LLL), representing the ability to continuously learn from experiences, is an essential characteristic of all living beings, enabling them to adapt and survive. However, learning and acquiring new knowledge from a series of tasks represents a challenge in artificial systems due to the catastrophic forgetting [1] which occurs when switching tasks.

Most existing studies address the forgetting from a series of predictive tasks, where the model is trained to remember the discriminative information across tasks. In this paper, we address a more challenging scenario in which the model is required to remember the generative representation information across domains over time. To implement this goal, we provide a mechanism for compressing and storing the probabilistic representations associated with the knowledge learnt from several data domains during LLL. For a given set of distinct domains (tasks) $\{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$, when learning the i-th task, we only access the data samples $\mathbf{x}$ drawn from a specific domain $\mathcal{T}_i$. Our goal is to find a model $\mathcal{M} = \{f_\omega, G_\varepsilon\}$ which can embed the knowledge from all prior tasks into a latent space $\mathcal{Z}$ through the inference process $f_\omega \colon \mathcal{X} \rightarrow \mathcal{Z}$ and recover the data from the embedded latent space $\mathcal{Z}$ through a generative process $G_\varepsilon \colon \mathcal{Z} \rightarrow \mathcal{X}$. Once $\mathcal{M}$ was trained, we can easily implement many down-stream tasks on the embedded latent space such as interpolation [2, 3] and log-likelihood estimation [4, 5]. This learning process opens a new direction for LLL where the model compresses the accumulated knowledge from a sequence of tasks into a compact latent space.

The primary challenge when attempting to learn multiple tasks is the catastrophic forgetting. Many approaches aiming to alleviate catastrophic forgetting are based on episodic memory systems [6] or by using Generative Replay Mechanisms (GRMs) [7]. In this paper, we focus on the GRM based models, since such methods do not rely on real data from prior tasks. Existing GRM models, although successfully used for LLL, fail to learn long sequences of tasks, where each database is characterized by different probabilistic representations. This is due to the mode collapse [8], when GRMs learn several entirely different data. It was shown that by employing expansible mixture models we can deal with such challenges. However, existing mixture models [4, 9, 5, 10] require preserving the whole network architecture while performing the model selection at the testing phase.

Inspired by addressing the drawbacks of GRMs while employing mixture models, we propose a new Lifelong Learning Framework (LKD), consisting of a Teacher-Student model, where the Teacher evolves over time according to the expansion mechanism to accumulate knowledge from a dynamically changing environment. The Student is designed to continually embed generative factors from the knowledge learned from the Teacher into a single latent space in which different data domains are embedded into multiple clusters. To ensure a compact network architecture for the Teacher, we propose to calculate the dependency between the incoming task and each expert through a knowledge consistency evaluation approach, which guides the selection and expansion of the experts in the Teacher. The main contributions are : (1) We propose

[*] Dr. A. G. Bors acknowledges the partial support from the EPSRC, UK, project COUSIN (EP/V009591/1)

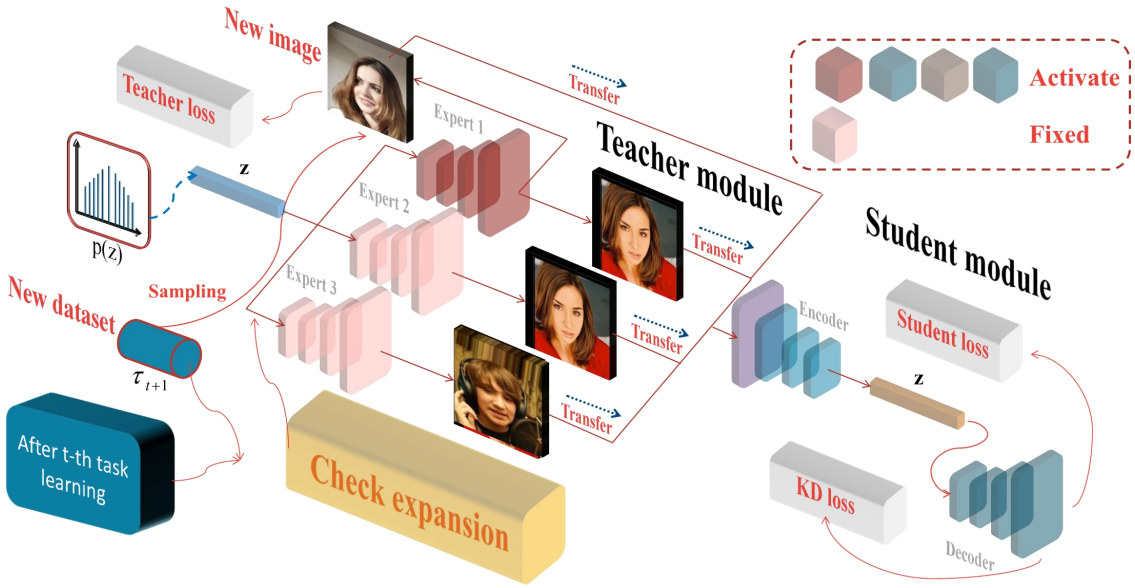


Fig. 1. The Lifelong Knowledge Distillation (LKD) framework consisting of the Teacher and Student modules.

to learn the accumulated generative factors, from successive domains by developing a novel lifelong learning framework; (2) A new approach to regularize the selection and expansion of the Teacher, which ensures a compact network architecture during training; (3) We introduce a new knowledge distillation loss that can distill the generative representations from the Teacher to the Student.

2. LIFELONG KNOWLEDGE DISTILLATION FRAMEWORK

2.1. Preliminaries

Problem definition: Let $\{\mathcal{D}_1^T, \cdots, \mathcal{D}_N^T\}$ be the training sets of N tasks, where each $\mathcal{D}_i^T$ consists of N_i^T testing samples $\mathbf{x}_i^T$ over the data space $\mathcal{X}$. This paper focuses on the cross-domain setting, aiming to learn a sequence of several datasets. Let $\mathcal{T} = \{\mathcal{T}_1, \cdots, \mathcal{T}_N\}$ be a set of N tasks, where each task $\mathcal{T}_i$ is defined by a training set $\mathcal{D}_i^S$. When learning $\mathcal{T}_i$, during the lifelong learning, a model would draw samples from the training set $\mathcal{D}_i^S$, while all previous training sets $\{\mathcal{D}_1^S, \cdots, \mathcal{D}_{i-1}^S\}$ are not available. Once the learning of all tasks is finished, the model's performance is evaluated on $\{\mathcal{D}_1^T, \cdots, \mathcal{D}_N^T\}$.

Variational Autoencoder (VAE) is an explicit generative latent variable model which learns an observed variable $\mathbf{x}$, while estimating a latent variable $\mathbf{z}$ over the latent space $\mathcal{Z}$ within a unified optimization framework. For a given VAE model $p_\theta(\mathbf{x}, \mathbf{z})$, we aim to search the optimal parameters θ that maximize the marginal log-likelihood $\log p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z}$. This involves the prior distribution $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$ (Gaussian distribution with an unit vector $\mathbf{I}$), which is intractable to optimize since it requires access to all $\mathbf{z}$. VAEs training relies on maximizing an Evidence Lower Bound (ELBO) to the marginal log-likelihood by, [11] :

$$\begin{aligned} \mathcal{L}_{ELBO}(\mathbf{x}; \theta, \xi) :=& \mathbb{E}_{z \sim q_\xi(\mathbf{z} \,|\, \mathbf{x})} \left[\log p_\theta(\mathbf{x} \,|\, \mathbf{z}) \right] \\ & - D_{KL} \left[q_\xi(\mathbf{z} \,|\, \mathbf{x}) \,||\, p(\mathbf{z}) \right] , \end{aligned} \tag{1}$$

where $q_\xi(\mathbf{z} \,|\, \mathbf{x})$ is a variational distribution aiming to approximate the true posterior $p_\theta(\mathbf{z} \,|\, \mathbf{x})$. $p_\theta(\mathbf{x} \,|\, \mathbf{z})$ is the decoding distribution and $D_{KL}(\cdot)$ represents the Kullback–Leibler (KL) divergence.

2.2. Teacher module

Using a single VAE for learning frequent GRM processes has significant limitations when learning a sequence of tasks [2]. In this paper, we develop a novel infinite mixture of VAEs (experts) as a dynamically expandable experts-based memory system for the Teacher module, where each expert captures one or several similar visual concepts from several given tasks. Let us assume that we have already trained K experts $\mathbf{M} = \{\mathcal{M}_1, \cdots, \mathcal{M}_K\}$ after $\mathcal{T}_t$, and each component $\mathcal{M}_i$ has the parameter set $\{\theta_i, \xi_i\}$. The dynamic expansion and selection mechanism is shown in Fig. 1. We require to evaluate the knowledge similarity between the information accumulated by each expert and that corresponding to the incoming task, in order to guide the Teacher module to adopt the appropriate learning strategy for the next task $\mathcal{T}_{t+1}$. In the following, we describe how we perform the selection and expansion for the next task learning.

Selection and expansion. As shown in Fig. 1. once the $\mathcal{T}_t$-th task was learnt, the selection or expansion procedure is performed by a non-parametric inference process in which we firstly evaluate the probability r, for mixture's expansion or component selection, by comparing the knowledge measure

$\min\{\mathrm{F}_{ks}(\mathcal{M}_i, \mathcal{T}_{t+1})\}_{i=1}^{K}$ and a threshold $hold$:

$$r = \begin{cases} 0, & \min\left\{\mathrm{F}_{ks}\left(\mathcal{M}_i, \mathcal{T}_{t+1}\right)\right\}_{i=1,\ldots,K} > hold\,; \\ 1, & \min\left\{\mathrm{F}_{ks}\left(\mathcal{M}_i, \mathcal{T}_{t+1}\right)\right\}_{i=1,\ldots,K} \leq hold\,, \end{cases} \tag{2}$$

where $\mathrm{F}_{ks}(\cdot)$ is a pre-defined function that evaluates the knowledge similarity. Then we update the selection probability p_i of each expert as :

$$p_i = \begin{cases} \dfrac{r \times \left(1/\mathrm{F}_{ks}\left(\mathcal{M}_i, \mathcal{T}_{t+1}\right)\right)}{\sum_{j=1}^{K}\left(1/\mathrm{F}_{ks}\left(\mathcal{M}_j, \mathcal{T}_{t+1}\right)\right)}, & i < K+1\,; \\ 1-r, & i = K+1\,. \end{cases} \tag{3}$$

If $r = 0$, then the Teacher module expands its capacity, $p_{(K+1)} = 1$, otherwise the Teacher module selects an expert according to $\{p_1, \ldots, p_K\}$.

Training the infinite mixture model. After determining the selection probability, we define the Teacher's loss function for the following $(t+1)$-th task, as :

$$\max_{\Theta} \sum\nolimits_{i=1}^{S^*} w_i \Big\{ \mathbb{E}_{z\sim q_\xi(\mathbf{z}\,|\,\mathbf{x})}\left[\log p_\theta(\mathbf{x}\,|\,\mathbf{z})\right] - D_{KL}\left[q_\xi(\mathbf{z}\,|\,\mathbf{x})\,||\,p(\mathbf{z})\right] \Big\}\,, \tag{4}$$

where $S^\star$ is the potential number of experts, determined by $S^* = K$ if the Teacher does not expand ($r = 1$) at the $(t+1)$-th task learning, otherwise a new expert is added, $S^\star = K+1$. Θ is the set of all components parameters. Then we train the Teacher by using Eq. (4) with the expert's weights $\mathbf{w} = \{w_1, \cdots, w_{S^\star}\}$ sampled from a Categorical distribution $\sim Cat(p_1, \ldots, p_{S^\star})$, optimizing either a selected, or a newly created component at $\mathcal{T}_{t+1}$.

2.3. Knowledge similarity (KS) evaluation

Existing mixture models [13, 14] use the log-likelihood, evaluated by each component when provided with a new training set, for the expansion or selection process. However, these models have the following drawbacks: 1) They require the existence of inference mechanisms for each component; 2) They do not have a mechanism to compare the statistical representations of a given task to their components' representations. In the following we address these shortcomings by proposing two approaches for evaluating the KS between each expert and the incoming task, without requiring any inference mechanism for the experts.

Firstly, we assume that we have the Student module, implemented by a single VAE, which is trained to learn the knowledge from all experts. The Student module already knows the whole information learnt so far and we can use its representation for the KS evaluation. We employ the cosine distance on the feature space $\mathcal{Z}$ of the Student for the KS evaluation :

$$\mathrm{COS}(\mathcal{M}_i, \mathcal{T}_{t+1}) := \frac{1}{n}\sum\nolimits_{u=1}^{m}\left\{\frac{\mathbf{z}_{t+1,u}\cdot\mathbf{z}'_{j,u}}{\|\mathbf{z}_{t+1,u}\|\,\|\mathbf{z}'_{j,u}\|}\right\}\,, \tag{5}$$

where $\mathbf{z}_{t+1,u}$ and $\mathbf{z}'_{j,u}$ are the feature vectors extracted by the Student when considering the inputs $\mathbf{x}_{t+1,u}$ and $\mathbf{x}'_{j,u}$, drawn from the incoming task $\mathcal{T}_{t+1}$ and the j-th expert, respectively. m is the number of samples used for the evaluation. Given that a larger measure in Eq. (5) represents a better similarity, we consider the following KS criterion in Eq (2) :

$$\mathrm{F}_{ks}\left(\mathcal{M}_i, \mathcal{T}_{t+1}\right)) = -\mathrm{COS}(\mathcal{M}_i, \mathcal{T}_{t+1}). \tag{6}$$

Secondly, we introduce another approach to evaluate the KS by comparing the sample log-likelihood between the incoming task and each expert, estimated using the Student model with parameters $\{\theta_s, \xi_s\}$:

$$\mathrm{F}_{Log}(\mathcal{M}_i, \mathcal{T}_{t+1}) := \frac{1}{n}\sum\nolimits_{u=1}^{m}\Big\{|\mathcal{L}_{ELBO}(\mathbf{x}_{t+1,u};\theta_s,\xi_s) - \mathcal{L}_{ELBO}(\mathbf{x}'_{j,u};\theta_s,\xi_s)|\Big\}\,, \tag{7}$$

where F_{ks} in Eq. (2) is implemented by F_{Log} in this case.

2.4. Data-free knowledge distillation (KD)

Unlike in other KD approaches that transfer knowledge at the predictive tasks [15], the proposed KD transfers data representation information through a sampling procedure without accessing real samples and labels. In order to embed the information from all experts into a single latent space, we implement the Student module as a VAE of parameters $\{\theta_s, \xi_s\}$. In the following, we introduce a new KD-based loss function that encourages the knowledge transfer on the posterior and the decoding distribution between the Teacher and Student:

$$\mathcal{L}_{KD1} = \sum\nolimits_{i=1}^{K} D_{KL}(p_{\theta_i}(\mathbf{x}\,|\,\mathbf{z})\,||\,p_{\theta_s}(\mathbf{x}\,|\,\mathbf{z})) \tag{8}$$

$$\mathcal{L}_{KD2} = \sum\nolimits_{i=1}^{K} D_{KL}(q_{\xi_i}(\mathbf{z}\,|\,\mathbf{x})\,||\,q_{\xi_s}(\mathbf{z}\,|\,\mathbf{x}))\,, \tag{9}$$

where $p_{\theta_s}(\mathbf{x}\,|\,\mathbf{z})$ and $q_{\xi_s}(\mathbf{z}\,|\,\mathbf{x})$ are the decoding and variational distributions for the Student module. Since we can not access past samples when evaluating Eq. (8) and (9), we estimate the KL divergence using the sampling process, where the past data $\mathbf{x}$ is generated by each expert and is then used as input for the inference model of the same expert. Together with the KD process, we introduce a new objective function allowing the Student to learn novel knowledge without forgetting previously learnt information :

$$\mathcal{L}_{Stu} = -\,\mathcal{L}_{ELBO}(\mathbf{x};\theta_s,\xi_s) + \eta_1\mathcal{L}_{KD1} + \eta_2\mathcal{L}_{KD2}\,, \tag{10}$$

where $\{\eta_1, \eta_2\}$ are hyperparameters balancing the learning of a new task and the already learnt knowledge. In practice, we divide the optimization of Eq. (10) into two independent optimization processes, where one is used to learn the new task only by minimizing $-\mathcal{L}_{ELBO}(\mathbf{x};\theta_s,\xi_s)$ and the second one is used for the knowledge transfer by minimizing $\mathcal{L}_{KD1} + \eta\mathcal{L}_{KD2}$. We set $\eta = 0.001$ in all experiments.

	SL					SSMI					PSNR				
Datasets	LGM	LKD	BE-Stu	LTS	LIMix-Stu	LGM	LKD	BE-Stu	LTS	LIMix-Stu	LGM	LKD	BE-Stu	LTS	LIMix-Stu
MNIST	51.15	50.53	33.65	75.40	176.82	0.81	0.81	0.86	0.71	0.42	18.34	18.18	20.16	16.56	13.72
Fashion	289.63	39.09	234.66	46.53	178.04	0.40	0.66	0.41	0.71	0.37	10.63	14.39	12.33	17.76	8.81
SVHN	309.66	33.85	113.00	63.99	146.70	0.25	0.78	0.45	0.45	0.47	7.56	19.35	11.01	11.03	13.58
IFashion	263.02	52.06	100.97	39.81	158.18	0.30	0.69	0.60	0.75	0.43	7.27	15.24	15.46	18.05	14.17
RMNIST	21.53	22.17	24.03	25.45	157.55	0.91	0.91	0.90	0.89	0.43	22.01	21.94	21.58	21.31	14.18
Average	187.00	**39.54**	101.26	50.24	163.45	0.53	**0.77**	0.64	0.70	0.42	13.16	**17.82**	16.11	16.94	12.89

Table 1. The performance of various models under the MSFIR setting, where the result for LIMix-Stu is reported from [12].

Dataset	LKD	LGM	BE-Stu	LTS
CelebA	63.26	740.36	138.22	243.75
CACD	138.51	1084.60	243.45	294.29
CIFAR10	235.55	590.56	268.95	224.52
Sub-ImageNet	241.36	624.86	271.08	232.29
SVHN	26.48	52.80	59.29	44.15
MNIST	28.41	21.39	26.02	25.49
Average	**138.93**	519.09	167.83	177.42

Table 2. Average Squared Loss (SL) when learning datasets with complex images.

Eq. (9) encourages the knowledge transfer on the generative representations from each expert to the inference model of the Student, while Eq. (8) ensures the consistency on the decoding distributions between the Student and Teacher modules.

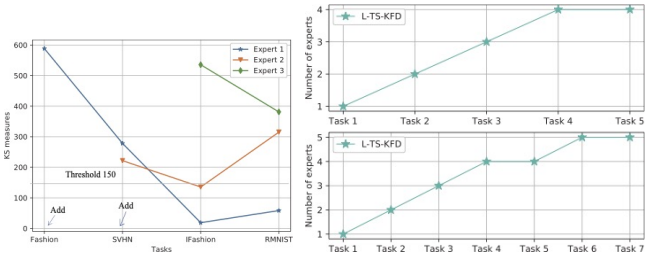
3. EXPERIMENTS

3.1. Datasets and evaluation criteria

Baselines. We compare the proposed Lifelong Knowledge Distillation (LKD) with several baselines including LTS [16], LGM [17] and BE-Stu. We implement BE-Stu by using the BatchEnsemble [18] as the Teacher, where each component is a VAE model and shares parameters between components. The Student in BE-Stu is implemented by a VAE model which is trained on the data generations by the Teacher. We also compare with LIMix-Stu [12] which uses the Teacher-Student framework. In our experiments, we mainly compare the LKD with the expansion and selection criterion from Eq. (7). We name our model as LKD-COS when the selection criterion is given by Eq. (5). We employ the Squared Loss (SL), the structural similarity index measure (SSIM) [19] and the Peak-Signal-to-Noise Ratio (PSNR) [19] as performance criteria.
Datasets. We follow the lifelong training setting from [12] which considers a sequence of five tasks, MNIST , SVHN, Fashion, InverseFashion (IFashion) and Rotated MNIST (RMNIST). We name this learning setting as MSFIR. We also consider a sequence of datasets containing complex images including the CelebA, CACD, CIFAR10, Sub-ImageNet, SVHN and MNIST, namely (CCCSSM).

3.2. The evaluation of the Student's performance

Firstly, we train all models under the MSFIR lifelong learning, where LKD uses three experts and the results from Table 1 show that LKD outperforms all baselines in every task under three criteria.



(a) KS for LKD training. (b) Network expansion in LKD.

Fig. 2. Knowledge similarity (KS) evaluation and the expansion of the network during the training.

We investigate the expansion of the LKD during LLL by evaluating the Knowledge Similarity (KS) measures $\mathrm{F}_{log}(\mathcal{M}_i, \mathcal{T}_{t+1})$ from Eq. (7) after each task switch, and the results are shown in Fig. 2-a. After learning the first task, the KS measure between the first expert and the data representation for the next task (SVHN database), is 586. Therefore, the Teacher module adds a new generator to learn SVHN. Then, after learning the third task, the KS measure between each expert and the next task (IFashion database) is smaller than 150. Therefore the Teacher module reuses the first expert to learn IFashion and RMNIST (last task). The architecture expansion of the Teacher module is shown in Fig. 2-b, where LKD leads to a reasonable number of experts, each capturing different data representations from the databases.

We also investigate the model's performance on databases with complex images. We train various models under (CCCSSM) setting where LKD uses $hold = 120$ in Eq. (2). The average SL results are provided in Table 2 indicating that the proposed framework performs better than other methods by a large margin. The performance of LTS tends to degenerate on CCCSSM since a single GAN is struggling to learn several complex databases through the GRM process. The expansion of the Teacher is shown in Fig. 2-b, where LKD uses 5 experts.

4. CONCLUSION

This research study is the first to explore Knowledge Distillation for transferring generative factors from multiple domains under the LLL framework. We achieve this by proposing an expanding model for the Teacher module within a Teacher-Student framework. We perform several experiments, and the empirical results verify the effectiveness and performance of the proposed methodology.

5. 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] Fei Ye and Adrian G. Bors, "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.

[3] Fei Ye and Adrian 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.

[4] Fei Ye and Adrian G. Bors, "Deep mixture generative autoencoders," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5789–5803, 2022.

[5] Fei Ye and Adrian G Bors, "Mixtures of variational autoencoders," in *Proc. Int. Conf. on Image Processing Theory, Tools and Applications (IPTA)*, 2020, pp. 1–6.

[6] Fei Ye and Adrian G. Bors, "Continual variational autoencoder learning via online cooperative memorization," in *Proc. European Conf. on Computer Vision (ECCV), vol. LNCS 13683*, 2022, pp. 531–549.

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

[8] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "Veegan: Reducing mode collapse in gans using implicit variational learning," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 2017, pp. 3308–3318.

[9] Fei Ye and Adrian G. Bors, "Lifelong mixture of variational autoencoders," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 461–474, 2023.

[10] Fei Ye and Adrian G Bors, "Lifelong generative modelling using dynamic expansion graph model," in *AAAI on Artificial Intelligence*, 2022, pp. 8857–8865.

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

[12] Fei Ye and Adrian G. Bors, "Lifelong infinite mixture model based on knowledge-driven Dirichlet process," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10695–10704.

[13] Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim, "A neural Dirichlet process mixture model for task-free continual learning," in *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:2001.00689*, 2020.

[14] Dushyant Rao, Francesco Visin, Andrei A. Rusu, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell, "Continual unsupervised representation learning," in *Advances Neural Information Processing Systems (NeurIPS)*, 2019, pp. 7647–7657.

[15] Yoon Kim and Alexander M Rush, "Sequence-level knowledge distillation," in *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, 2016, pp. 1317–1327.

[16] Fei Ye and Adrian G. Bors, "Lifelong teacher-student network learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6280–629, 2022.

[17] Jason Ramapuram, Magda Gregorova, and Alexandros Kalousis, "Lifelong generative modeling," *Neurocomputing*, vol. 404, pp. 381–400, 2020.

[18] Yeming Wen, Dustin Tran, and Jimmy 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.

[19] Alain Hore and Djemel Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. International Conference on Pattern Recognition (ICPR)*, 2010, pp. 2366–2369.