Continual Unsupervised Generative Modeling Supplementary Materials

Fei Ye and Adrian G. Bors

APPENDIX A
THE PROOF OF THEOREM 1

Assumption 1. Let $\mathcal{X}$ be a metric space that satisfies $\mathcal{L}(a,b) \leq \mathcal{L}(a,c) + \mathcal{L}(c,b)$ where the loss function $\mathcal{L}(\cdot)$ is a metric and $a, b, c \in \mathcal{X}$.

Based on Assumption 1, we provide the detailed proof as follows :

Proof. Let $\mathcal{P}_i$ and $\tilde{\mathcal{P}}_i$ be two domains over $\mathcal{X}$. Then for $h^{\star}_{\mathcal{P}_i} = \arg\min_{h\in\mathcal{H}} \mathcal{E}_{\mathcal{P}_i}(h, f_{\mathcal{P}_i})$ and $h^{\star}_{\tilde{\mathcal{P}}_i} = \arg\min_{h\in\mathcal{H}} \mathcal{E}_{\tilde{\mathcal{P}}_i}(h, f_{\tilde{\mathcal{P}}_i})$ where $f_{\tilde{\mathcal{P}}_i} \in \mathcal{H}$ is the ground truth function (identity function under the encoder-decoding process) for $\tilde{\mathcal{P}}_i$.

Then according to the triangle inequality property of $\mathcal{L}$, applied twice, we have :

$$\begin{aligned}\mathcal{E}_{\mathcal{P}_i}(h, f_{\mathcal{P}_i}) \leq \mathcal{E}_{\mathcal{P}_i}(h, h^*_{\tilde{\mathcal{P}}_i}) + \mathcal{E}_{\mathcal{P}_i}(h^*_{\tilde{\mathcal{P}}_i}, h^*_{\mathcal{P}_i}) \\ + \mathcal{E}_{\mathcal{P}_i}(h^*_{\mathcal{P}_i}, f_{\mathcal{P}_i})\end{aligned} \tag{1}$$

Eq. (1) holds because, after applying twice the triangle inequality, $\mathcal{L}(a,b) \leq \mathcal{L}(a,c)+\mathcal{L}(c,d)+L(d,b)$ where a, b, c, d are $h(\mathbf{x}), f_{\mathcal{P}_i}(\mathbf{x}), h^*_{\tilde{\mathcal{P}}_i}(\mathbf{x}), h^*_{\mathcal{P}_i}(\mathbf{x})$ and $\mathbf{x}$ is sampled from the same domain $\mathcal{P}_i$. According to the definition of discrepancy distance (See Definition 2 from the paper), we define the discrepancy distance between $\mathcal{P}_i$ and $\tilde{\mathcal{P}}_i$ as :

$$\begin{aligned}\mathcal{L}_{\text{disc}}(\mathcal{P}_i, \tilde{\mathcal{P}}_i) = \sup_{(h,h')\in\mathcal{H}} |\mathbb{E}_{\mathbf{x}\sim\mathcal{P}_i}\left[\mathcal{L}\left(h'(\mathbf{x}), h(\mathbf{x})\right)\right] \\ - \mathbb{E}_{\mathbf{x}\sim\tilde{\mathcal{P}}_i}\left[\mathcal{L}\left(h'(\mathbf{x}), h(\mathbf{x})\right)\right]|.\end{aligned} \tag{2}$$

We rewrite the above equation as :

$$\mathcal{L}_{\text{disc}}(\mathcal{P}_i, \tilde{\mathcal{P}}_i) = \sup_{(h,h')\in\mathcal{H}} \left|\mathcal{E}_{\mathcal{P}_i}(h,h') - \mathcal{E}_{\tilde{\mathcal{P}}_i}(h,h')\right|. \tag{3}$$

We consider h' to be $h^*_{\tilde{\mathcal{P}}_i}$ in Eq. (3) and we have :

$$\begin{aligned}\sup_{(h,h')\in H} \left|\mathcal{E}_{\mathcal{P}_i}(h,h') - \mathcal{E}_{\tilde{\mathcal{P}}_i}(h,h')\right| \geq \\ \left|\mathcal{E}_{\mathcal{P}_i}(h, h^*_{\tilde{\mathcal{P}}_i}) - \mathcal{E}_{\tilde{\mathcal{P}}_i}(h, h^*_{\tilde{\mathcal{P}}_i})\right|\end{aligned} \tag{4}$$

We also know that :

$$\begin{aligned}\mathcal{E}_{\mathcal{P}_i}(h, h^*_{\tilde{\mathcal{P}}_i}) \leq \left|\mathcal{E}_{\mathcal{P}_i}(h, h^*_{\tilde{\mathcal{P}}_i}) - \mathcal{E}_{\tilde{\mathcal{P}}_i}(h, h^*_{\tilde{\mathcal{P}}_i})\right| \\ + \mathcal{E}_{\tilde{\mathcal{P}}_i}(h, h^*_{\tilde{\mathcal{P}}_i})\end{aligned} \tag{5}$$

Therefore, we can replace the first term of the right hand side of Eq. (1) by the right hand side of Eq. (5), resulting in :

$$\begin{aligned}\mathcal{E}_{\mathcal{P}_i}(h, f_{\mathcal{P}_i}) \leq \mathcal{E}_{\tilde{\mathcal{P}}_i}(h, h^*_{\tilde{\mathcal{P}}_i}) \\ + \left|\mathcal{E}_{\mathcal{P}_i}(h, h^*_{\tilde{\mathcal{P}}_i}) - \mathcal{E}_{\tilde{\mathcal{P}}_i}(h, h^*_{\tilde{\mathcal{P}}_i})\right| \\ + \mathcal{E}_{\mathcal{P}_i}(h^*_{\mathcal{P}_i}, h^*_{\tilde{\mathcal{P}}_i}) + \mathcal{E}_{\mathcal{P}_i}(h^*_{\mathcal{P}_i}, f_{\mathcal{P}_i})\end{aligned} \tag{6}$$

Then the second term, representing the absolute value of the difference in the RHS of Eq. (6) can be replaced by $\mathcal{L}_{\text{disc}}(\mathcal{P}_i, \tilde{\mathcal{P}}_i)$ from Eq. (4), since the discrepancy distance between two distributions is an upper bound to this absolute value, resulting in :

$$\begin{aligned}\mathcal{E}_{\mathcal{P}_i}(h, f_{\mathcal{P}_i}) \leq \mathcal{E}_{\tilde{\mathcal{P}}_i}(h, h^*_{\tilde{\mathcal{P}}_i}) + \mathcal{L}_{\text{disc}}(\mathcal{P}_i, \tilde{\mathcal{P}}_i) \\ + \mathcal{E}_{\mathcal{P}_i}(h^*_{\mathcal{P}_i}, h^*_{\tilde{\mathcal{P}}_i}) + \mathcal{E}_{\mathcal{P}_i}(h^*_{\mathcal{P}_i}, f_{\mathcal{P}_i})\end{aligned} \tag{7}$$

From **Definition 7** from the paper, we know that $\mathcal{L}_{\text{disc}}(\mathcal{P}_i, \tilde{\mathcal{P}}_i) \leq \mathcal{L}^{\star}_{\text{disc}}(\mathcal{P}_i, \tilde{\mathcal{P}}_i)$. We then replace $\mathcal{L}_{\text{disc}}(\mathcal{P}_i, \tilde{\mathcal{P}}_i)$ by using $\mathcal{L}^{\star}_{\text{disc}}(\mathcal{P}_i, \tilde{\mathcal{P}}_i)$ in Eq. (7), resulting in :

$$\begin{aligned}\mathcal{E}_{\mathcal{P}_i}(h, f_{\mathcal{P}_i}) \leq \mathcal{E}_{\tilde{\mathcal{P}}_i}(h, h^*_{\tilde{\mathcal{P}}_i}) + \mathcal{L}^{\star}_{\text{disc}}(\mathcal{P}_i, \tilde{\mathcal{P}}_i) \\ + \mathcal{E}_{\mathcal{P}_i}(h^*_{\mathcal{P}_i}, h^*_{\tilde{\mathcal{P}}_i}) + \mathcal{E}_{\mathcal{P}_i}(h^*_{\mathcal{P}_i}, f_{\mathcal{P}_i})\end{aligned} \tag{8}$$

Eq. (8) proves **Theorem 1** and a similar proof can be found in Theorem 8 from [1].

APPENDIX B
THE PROOF OF THEOREM 2

Theorem 2. For a given sequence of tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_t\}$, we derive a GB between the target distribution and the evolved source distribution during the t-th task learning :

$$\begin{aligned}\frac{1}{t}\sum\nolimits_{i=1}^{t} \mathcal{E}_{\mathcal{P}_i}(h, f_{\mathcal{P}_i}) \leq \mathcal{E}_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left(h, h^*_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\right) \\ + \mathcal{E}_R\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t\right),\end{aligned} \tag{9}$$

where $\mathcal{P}_{(1:t)}$ is the mixture distribution $\{\mathcal{P}_1 \otimes \mathcal{P}_2, \ldots, \otimes\mathcal{P}_t\}$. As it can be seen in Eq. (9), the performance on the target domain is largely depending on the discrepancy term even if $\mathcal{M}$ minimizes the source risk well. In the following, we provide an analytical bound that considers all previously learnt distributions.

$$\begin{aligned}\frac{1}{t}\sum\nolimits_{i=1}^{t} \mathcal{E}_{\mathcal{P}_i}(h, f_{\mathcal{P}_i}) \leq \mathcal{E}_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left(h, h^*_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\right) \\ + \text{Err}^a + \text{Err}^d,\end{aligned} \tag{10}$$

where $\mathrm{Err}^d \geq 0$ evaluates the difference on the two risk terms, expressed by:

$$\begin{aligned}&\sum_{k=1}^{t-1}\Big\{\mathcal{E}_{\mathbb{P}^{(t-k)}}\left(h, h^*_{\mathbb{P}^{(t-l)}}\right)\\&-\mathcal{E}_{\mathbb{P}^{(t-1-k)}\otimes\tilde{\mathcal{P}}_{(t-k)}}\left(h, h^*_{\mathbb{P}^{(t-1-k)}\otimes\tilde{\mathcal{P}}_{(t-k)}}\right)\Big\},\end{aligned} \tag{11}$$

where $\mathcal{E}_{\mathbb{P}^0\otimes\tilde{\mathcal{P}}_1}(h, h^*_{\mathbb{P}^0\otimes\tilde{\mathcal{P}}_1}) = \mathcal{E}_{\tilde{\mathcal{P}}_1}(h, h^*_{\tilde{\mathcal{P}}_1})$. Err^a is the accumulated error term expressed by :

$$\begin{aligned}&\sum_{k=1}^{t-2}\left\{\mathcal{E}_R\left(\mathbb{P}^{(t-1-k)}\otimes\tilde{\mathcal{P}}_{(t-k)}, \mathbb{P}^{(t-k)}\right)\right\}\\&+\mathcal{E}_R\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t\right)+\mathcal{E}_R\left(\tilde{\mathcal{P}}_1, \mathbb{P}^1\right).\end{aligned} \tag{12}$$

From Eq. (12), we can observe that while learning more tasks (t is increased) during the lifelong learning, the gap on the GB for $\mathcal{M}$ tends to become larger given that Err^a increases. This also explains why GR fails when learning a long sequence of tasks [2], [3]. Additionally, the term $\mathcal{E}_R(\mathbb{P}^{(t-1-k)}\otimes\tilde{\mathcal{P}}_{(t-k)}, \mathbb{P}^{(t-k)})$ and Err^d tend to be small when the discrepancy $\mathcal{L}^\star_{\mathrm{disc}}(\mathbb{P}^i, \mathbb{P}^{(i-1)}\otimes\tilde{\mathcal{P}}_i)$ is equal to 0 in each task learning ($i = 1, \ldots, t$). This is achieved by the optimal generator distribution that approximates $\mathbb{P}^{(i-1)}\otimes\tilde{\mathcal{P}}_i$ exactly in each task learning,

Proof. Firstly, we can derive the bound according to Theorem 1 :

$$\begin{aligned}\mathcal{E}_{\mathcal{P}_{(i:t)}}\left(h, f_{\mathcal{P}_{(i:t)}}\right) \leq\ &\mathcal{E}_{\mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t}\left(h, h^*_{\mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t}\right)\\&+\mathcal{L}^\star_{\mathrm{disc}}\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t\right)\\&+\varepsilon\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t\right)\end{aligned} \tag{13}$$

where $\mathcal{P}_{(1:t)}$ represents the mixture distribution $\{\mathcal{P}_1 \otimes \mathcal{P}_2, \ldots, \otimes\mathcal{P}_t\}$. Let $\rho_{(1:t)}(\mathbf{x})$ represent the density function for $\mathcal{P}_{(1:t)}$ and $\rho_{(i)}(\mathbf{x})$ the density function for $\mathcal{P}_i$. Since $\mathcal{P}_{(1:t)}$ is the mixture distribution and its density is expressed by $\rho_{(1:t)}(\mathbf{x}) = \frac{1}{t}\sum_{i=1}^t \rho_{(i)}(\mathbf{x})$. We know that $\mathcal{E}_{\mathcal{P}_{(i:t)}}(h, f_{\mathcal{P}_{(i:t)}})$ can be rewritten as the integral form $\int \rho_{(i:t)}(\mathbf{x})\mathcal{L}(h, f_{P_{(1:t)}})\,\mathrm{d}\mathbf{x}$. We then take $\rho_{(1:t)}(\mathbf{x}) = \frac{1}{t}\sum_{i=1}^t \rho_{(i)}(\mathbf{x})$ in this integral form, resulting in :

$$\frac{1}{t}\sum_{i=1}^t \int \rho_{(i)}(\mathbf{x})\mathcal{L}(h, f_{P_{(1:t)}})\mathrm{d}\mathbf{x} \tag{14}$$

We assume that $\mathcal{P}_i$ is independent from $\mathcal{P}_j$, where $i \neq j$, which is a reasonable assumption, since each task is associated with a different dataset. Therefore, the true labeling function $f_{\mathcal{P}_{(1:t)}}$ can be represented by $f_{\mathcal{P}_i}$ under the target distribution $\mathcal{P}_i$ of the i-th task. Then we rewrite Eq. (14) as the expectation form $\frac{1}{t}\sum_{i=1}^t \mathcal{E}_{\mathcal{P}_i}(h, f_{\mathcal{P}_i})$.

Based on the above results, Eq. (13) is rewritten as :

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^t \mathcal{E}_{\mathcal{P}_i}(h, f_{\mathcal{P}_i}) \leq\ &\mathcal{E}_{\mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t}\left(h, h^*_{\mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t}\right)\\&+\mathcal{L}^\star_{\mathrm{disc}}\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t\right)\\&+\varepsilon\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t\right)\end{aligned} \tag{15}$$

In the following, we provide the derivations for Err^a and Err^d from Eq. (10). We consider to take $\mathbb{P}^{t-2}\otimes\tilde{\mathcal{P}}_{t-1}$ and $\mathbb{P}^{t-1}$ as the target and source domains, respectively. This is a reasonable choice since we allow $\mathbb{P}^{t-1}$ (the generator distribution of the model) to approximate $\mathbb{P}^{t-2}\otimes\tilde{\mathcal{P}}_{t-1}$. We derive the bound as :

$$\begin{aligned}\mathcal{E}_{\mathbb{P}^{t-2}\otimes\tilde{\mathcal{P}}_{t-1}}\left(h, h^*_{\mathbb{P}^{t-2}\otimes\tilde{\mathcal{P}}_{t-1}}\right) \leq\ &\mathcal{E}_{\mathbb{P}^{t-1}}\left(h, h^*_{\mathbb{P}^{t-1}}\right)\\&+\mathcal{L}^\star_{\mathrm{disc}}\left(\mathbb{P}^{t-2}\otimes\tilde{\mathcal{P}}_{t-1}, \mathbb{P}^{t-1}\right)\\&+\varepsilon\left(\mathbb{P}^{t-2}\otimes\tilde{\mathcal{P}}_{t-1}, \mathbb{P}^{t-1}\right)\end{aligned} \tag{16}$$

We then consider to take $\mathbb{P}^{t-3}\otimes\tilde{\mathcal{P}}_{t-2}$ and $\mathbb{P}^{t-2}$ as the target and source domains, respectively, and we derive the bound as :

$$\begin{aligned}&\mathcal{E}_{\mathbb{P}^{t-3}\otimes\tilde{\mathcal{P}}_{t-2}}\left(h, h^*_{\mathbb{P}^{t-3}\otimes\tilde{\mathcal{P}}_{t-2}}\right) \leq \mathcal{E}_{\mathbb{P}^{t-2}}\left(h, h^*_{\mathbb{P}^{t-2}}\right)\\&+\mathcal{L}^\star_{\mathrm{disc}}\left(\mathbb{P}^{t-3}\otimes\tilde{\mathcal{P}}_{t-2}, \mathbb{P}^{t-2}\right)\\&+\varepsilon\left(\mathbb{P}^{t-3}\otimes\tilde{\mathcal{P}}_{t-2}, \mathbb{P}^{t-2}\right)\end{aligned} \tag{17}$$

According to the inductive inference, we have :

$$\begin{aligned}&\ldots\\&\ldots\\&\mathcal{E}_{\mathbb{P}^1\otimes\tilde{\mathcal{P}}_2}\left(h, h^*_{\mathbb{P}^1}\right) \leq \mathcal{E}_{\mathbb{P}^2}\left(h, h^*_{\mathbb{P}^2}\right)+\mathcal{L}^\star_{\mathrm{disc}}\left(\mathbb{P}^1\otimes\tilde{\mathcal{P}}_2, \mathbb{P}^2\right)\\&+\varepsilon\left(\mathbb{P}^1\otimes\tilde{\mathcal{P}}_2, \mathbb{P}^2\right)\end{aligned} \tag{18}$$

$$\begin{aligned}&\mathcal{E}_{\tilde{\mathcal{P}}_1}\left(h, h^*_{\tilde{\mathcal{P}}_1}\right) \leq \mathcal{E}_{\mathbb{P}^1}\left(h, h^*_{\mathbb{P}^1}\right)+\mathcal{L}^\star_{\mathrm{disc}}\left(\tilde{\mathcal{P}}_1, \mathbb{P}^1\right)\\&+\varepsilon\left(\tilde{\mathcal{P}}_1, \mathbb{P}^1\right)\end{aligned} \tag{19}$$

We then sum up all the above derivations in the inequality from Eq. (15) to Eq. (19), resulting in :

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^t\left\{\mathcal{E}_{\mathcal{P}_i}\left(h, f_{\mathcal{P}_i}\right)\right\}+\mathcal{E}_{\tilde{\mathcal{P}}_1}\left(h, h^*_{\tilde{\mathcal{P}}_1}\right)\\&+\sum_{k=1}^{t-2}\left\{\mathcal{E}_{\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}}\left(h, h^*_{\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}}\right)\right\} \leq\\&\sum_{k=1}^{t-2}\left\{\mathcal{E}_{\mathbb{P}^{t-k}}\left(h, h^*_{\mathbb{P}^{t-k}}\right)\right\}+\mathcal{E}_{\mathbb{P}^1}\left(h, h^*_{\mathbb{P}^1}\right)\\&+\mathcal{E}_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left(h, h^*_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\right)+\sum_{k=1}^{t-2}\Big\{\mathcal{L}^\star_{\mathrm{disc}}\left(\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}, \mathbb{P}^{t-k}\right)\\&+\varepsilon\left(\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}, \mathbb{P}^{t-k}\right)\Big\}\\&+\mathcal{L}^\star_{\mathrm{disc}}\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t\right)\\&+\varepsilon\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t\right)+\mathcal{L}^\star_{\mathrm{disc}}\left(\tilde{\mathcal{P}}_1, \mathbb{P}^1\right)+\varepsilon\left(\tilde{\mathcal{P}}_1, \mathbb{P}^1\right)\end{aligned} \tag{20}$$

Then we move the second and third term in the left hand side to the right hand side in Eq.(20), resulting in :

$$\begin{aligned}
&\frac{1}{t}\sum_{i=1}^{t}\left\{\mathcal{E}_{\mathcal{P}_i}\left(h, f_{\mathcal{P}_i}\right)\right\} \leq \mathcal{E}_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left(h, h^*_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\right) + \mathcal{E}_{\mathbb{P}^1}\left(h, h^*_{\mathbb{P}^1}\right) \\
&- \mathcal{E}_{\tilde{\mathcal{P}}_1}\left(h, h^*_{\tilde{\mathcal{P}}_1}\right) \\
&+ \sum_{k=1}^{t-2}\left\{\mathcal{E}_{\mathbb{P}^{t-k}}\left(h, h^*_{\mathbb{P}^{t-k}}\right) - \mathcal{E}_{\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}}\left(h, h^*_{\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}}\right)\right\} \\
&+ \sum_{k=1}^{t-2}\left(\mathcal{L}^{\star}_{\text{disc}}\left(\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}, \mathbb{P}^{t-k}\right)\right. \\
&\left.+\varepsilon\left(\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}, \mathbb{P}^{t-k}\right)\right) + \mathcal{L}^{\star}_{\text{disc}}\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t\right) \\
&+ \varepsilon\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t\right) + \mathcal{L}^{\star}_{\text{disc}}\left(\tilde{\mathcal{P}}_1, \mathbb{P}^1\right) + \varepsilon\left(\tilde{\mathcal{P}}_1, \mathbb{P}^1\right)
\end{aligned} \tag{21}$$

Then we can rewrite Eq. (21) as :

$$\begin{aligned}
&\frac{1}{t}\sum_{i=1}^{t}\left\{\mathcal{E}_{\mathcal{P}_i}\left(h, f_{\mathcal{P}_i}\right)\right\} \leq \mathcal{E}_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left(h, h^*_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\right) + \mathcal{E}_{\mathbb{P}^1}\left(h, h^*_{\mathbb{P}^1}\right) \\
&- \mathcal{E}_{\tilde{\mathcal{P}}_1}\left(h, h^*_{\tilde{\mathcal{P}}_1}\right) \\
&+ \underbrace{\sum_{k=1}^{t-1}\left\{\mathcal{E}_{\mathbb{P}^{t-k}}\left(h, h^*_{\mathbb{P}^{t-k}}\right) - \mathcal{E}_{\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}}\left(h, h^*_{\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}}\right)\right\}}_{\text{Err}^{\text{a}}} \\
&+ \sum_{k=1}^{t-2}\left\{\mathcal{E}_R\left(\mathbb{P}^{t-1-k}\otimes\tilde{\mathcal{P}}_{t-k}, \mathbb{P}^{t-k}\right)\right\} \\
&+ \mathcal{E}_R\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{(t-1)}\otimes\tilde{\mathcal{P}}_t\right) + \mathcal{E}_R\left(\tilde{\mathcal{P}}_1, \mathbb{P}^1\right)
\end{aligned} \tag{22}$$

where the latest three terms are Err^{d} and $\mathcal{E}_{\mathbb{P}^0\otimes\tilde{\mathcal{P}}_1}(h, h^*_{\mathbb{P}^0\otimes\tilde{\mathcal{P}}_1}) = \mathcal{E}_{\tilde{\mathcal{P}}_1}(h, h^*_{\tilde{\mathcal{P}}_1})$ and this proves Theorem 2.

APPENDIX C
THE PROOF OF LEMMA 1

According to the bound on the KL divergence :

$$\begin{aligned}
&\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_{\mathcal{P}_i} KL(q_{\omega^t}(\mathbf{z} \mid \mathbf{x}_i^T) \mid\mid p(\mathbf{z})) \leq \\
&\mathbb{E}_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t} KL(q_{\omega^t}(\mathbf{z} \mid \tilde{\mathbf{x}}^t) \mid\mid p(\mathbf{z})) \\
&+ \left|\mathbb{E}_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t} KL(q_{\omega^t}(\mathbf{z} \mid \tilde{\mathbf{x}}^t) \mid\mid p(\mathbf{z})) \right. \\
&\left. - \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_{\mathcal{P}_i} KL(q_{\omega^t}(\mathbf{z} \mid \mathbf{x}_i^T) \mid\mid p(\mathbf{z}))\right|
\end{aligned} \tag{23}$$

We also know that $\mathcal{L}_{ELBO}\left(\mathbf{x}; \{\theta, \omega\}\right)$ is expressed as :

$$\begin{aligned}
\mathcal{L}_{ELBO}\left(\mathbf{x}; \{\theta, \omega\}\right) := &\mathbb{E}_{q_\omega(\mathbf{z}|\mathbf{x})}\left[\log p_\theta\left(\mathbf{x} \mid \mathbf{z}\right)\right] \\
&- KL\left[q_\omega\left(\mathbf{z} \mid \mathbf{x}\right) \mid\mid p\left(\mathbf{z}\right)\right],
\end{aligned} \tag{24}$$

When the decoder models a Gaussian distribution, $\log p_\theta\left(\mathbf{x} \mid \mathbf{z}\right)$ can be represented as :

$$\log p_\theta\left(\mathbf{x} \mid \mathbf{z}\right) = -\frac{1}{2d\sigma_\theta^2\left(\mathbf{z}\right)}\left\|\mathbf{x} - \mu_\theta\left(\mathbf{z}\right)\right\|^2 - -\frac{1}{2}d\log\sqrt{2\pi}\sigma_\theta^2\left(\mathbf{z}\right) \tag{25}$$

where $\sigma_\theta\left(\mathbf{z}\right)$ and $\mu_\theta\left(\mathbf{z}\right)$ are the variance and standard deviation of the Gaussian distribution, obtained by the decoder while d is the dimension. $\|\cdot\|^2$ represents the reconstruction error (square loss). We implement the decoder by a Gaussian distribution with the identical standard deviation for all dimensions, $\mathcal{N}(\mu_\theta\left(\mathbf{z}\right), \sigma\mathbf{I})$ where $\mu_\theta\left(\mathbf{z}\right)$ is modeled by a deep convolutional neural network and $\mathbf{I}$ is the identity matrix. Therefore, Eq. (25) is represented by the fixed standard deviation σ :

$$\log p_\theta\left(\mathbf{x} \mid \mathbf{z}\right) = -\frac{1}{2\sigma^2}\left\|\mathbf{x} - \mu_\theta\left(\mathbf{z}\right)\right\|^2 - -\frac{1}{2}d\log\sqrt{2\pi}\sigma^2 \tag{26}$$

Since h is the hypothesis of the model ($\mathcal{M}^t$), implemented as a encoding-decoding process, we have

$$\begin{aligned}
\mathcal{L}_{ELBO}(\mathbf{x}_i^T; h) = &-\frac{1}{2\sigma^2}\mathcal{L}(h(\mathbf{x}_i^T), f_{\mathcal{P}_i}(\mathbf{x}_i^T)) - \frac{1}{2}d\log\sqrt{2\pi}\sigma^2 \\
&- KL(q_{\omega^t}(\mathbf{z} \mid \mathbf{x}_i^T) \mid\mid p(\mathbf{z}))
\end{aligned} \tag{27}$$

Then we evaluate the negative ELBO :

$$\begin{aligned}
-\mathcal{L}_{ELBO}(\mathbf{x}_i^T; h) = &\frac{1}{2\sigma^2}\mathcal{L}(h(\mathbf{x}_i^T), f_{\mathcal{P}_i}(\mathbf{x}_i^T)) + \frac{1}{2}d\log\sqrt{2\pi}\sigma^2 \\
&+ KL(q_{\omega^t}(\mathbf{z} \mid \mathbf{x}_i^T) \mid\mid p(\mathbf{z})).
\end{aligned} \tag{28}$$

And we also know that :

$$\mathcal{R}_{\mathcal{P}_i}(h, f_{\mathcal{P}_i}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_i}\mathcal{L}(h(\mathbf{x}_i^T), f_{\mathcal{P}_i}(\mathbf{x}_i^T)) \tag{29}$$

and we have :

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim\mathcal{P}_i}\left[-\mathcal{L}_{ELBO}(\mathbf{x}_i^T; h)\right] = &\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\{\frac{1}{2\sigma^2}\mathcal{L}(h(\mathbf{x}_i^T), f_{\mathcal{P}_i}(\mathbf{x}_i^T)) \\
&+ KL(q_{\omega^t}(\mathbf{z} \mid \mathbf{x}_i^T) \mid\mid p(\mathbf{z}))\} \\
&+ \frac{1}{2}d\log\sqrt{2\pi}\sigma^2
\end{aligned} \tag{30}$$

We observe that $\frac{1}{2}d\log\sqrt{2\pi}\sigma^2$ and $\frac{1}{2d\sigma^2}$ are constants. In order to simplify the notations, we set $\sigma = \frac{1}{\sqrt{2}\sqrt{d}}$. Therefore, Eq. (30) is rewritten as :

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\left[-\mathcal{L}_{ELBO}(\mathbf{x}_i^T; h)\right] = &\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\{\mathcal{L}(h(\mathbf{x}_i^T), f_{\mathcal{P}_i}(\mathbf{x}_i^T)) \\
&+ KL(q_{\omega^t}(\mathbf{z} \mid \mathbf{x}_i^T) \mid\mid p(\mathbf{z}))\} \\
&+ \frac{1}{2}d\log\sqrt{2\pi}\frac{1}{2d}
\end{aligned} \tag{31}$$

Eq. (31) can be seen as the average ELBO for all samples. We then take Eq. (23) into Eq. (7) from the paper, and we have :

$$\begin{aligned}
&\frac{1}{t}\sum_{i=1}^{t}\left\{\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\{\mathcal{L}(h(\mathbf{x}_i^T), f_{\mathcal{P}_i}(\mathbf{x}_i^T)) + KL(q_{\omega^t}(\mathbf{z}|\mathbf{x}_i^T) \mid\mid p(\mathbf{z}))\}\right\} \leq \\
&\mathbb{E}_{\mathbf{x}^t\sim\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left\{\mathcal{L}\left(h(\tilde{\mathbf{x}}^t), h^*_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}(\tilde{\mathbf{x}}^t)\right)\right. \\
&\left. + KL(q_{\omega^t}(\mathbf{z} \mid \tilde{\mathbf{x}}^t) \mid\mid p(\mathbf{z}))\right\} \\
&+ |KL_1 - KL_2| \; + \mathcal{E}_R\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t\right)
\end{aligned} \tag{32}$$

It notes that we can add the constant $\frac{1}{2}\log\pi$ in both sides of Eq. (32). According to Eq. (31), we can rewrite Eq. (32) as :

$$\begin{aligned}
&\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}[-\mathcal{L}_{ELBO}\left(\mathbf{x}_i^T; h\right)] \leq \\
&\mathbb{E}_{\mathbf{x}^t\sim\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left[-\mathcal{L}_{ELBO}\left(\tilde{\mathbf{x}}^t; h\right)\right] + |KL_1 - KL_2| \\
&+ \mathcal{E}_R\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t\right).
\end{aligned} \tag{33}$$

This proves Lemma 1. We can further replace the last term from the right hand side (RHS) of Eq. (33) by $\text{Err}^a + \text{Err}^d$ (See details in the proof of Theorem 2), resulting in :

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}[-\mathcal{L}_{ELBO}\left(\mathbf{x}_i^T;h\right)]\leq\\&\mathbb{E}_{\mathbf{x}^t\sim\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left[-\mathcal{L}_{ELBO}\left(\tilde{\mathbf{x}}^t;h\right)\right]+|KL_1-KL_2|\\&+\text{Err}^a+\text{Err}^d.\end{aligned}\tag{34}$$

APPENDIX D
THE PROOF OF THEOREM 3

Firstly, $\mathcal{E}_C$ can be easily proved since the task was trained only once and we simply derive the bound between the training sets and testing sets.

For the components that are trained more than once, we firstly consider the c_i'-th component and $a(i,j)$-th task. This can be easily generalized to other components and other tasks. We firstly consider to take $\mathcal{P}_{a(i,j)}$ and $\mathbb{P}^0_{a(i,j)}$ as the target and source distribution and we have a bound :

$$\begin{aligned}&\mathcal{E}_{\mathcal{P}_{a(i,j)}}\left(h_{c_i'},f_{\mathcal{P}_{a(i,j)}}\right)\leq\mathcal{E}_{\mathbb{P}^0_{a(i,j)}}\left(h_{c_i'},h^*_{\mathbb{P}^0_{a(i,j)}}\right)\\&+disc^\star_{\mathcal{L}}\left(\mathcal{P}_{a(i,j)},\mathbb{P}^0_{a(i,j)}\right)+\varepsilon\left(\mathcal{P}_{a(i,j)},\mathbb{P}^0_{a(i,j)}\right)\end{aligned}\tag{35}$$

We can observe that $\mathbb{P}^0_{a(i,1)}$ represent the training set $\tilde{\mathcal{P}}_{a(i,j)}$. Then we consider to take $\mathbb{P}^0_{a(i,j)}$ and $\mathbb{P}^1_{a(i,j)}$ as the target and source domains, respectively. Then we have the bound as :

$$\begin{aligned}&\mathcal{E}_{\mathbb{P}^0_{a(i,j)}}\left(h_{c_i'},f_{\mathbb{P}^0_{a(i,j)}}\right)\leq\mathcal{E}_{\mathbb{P}^1_{a(i,j)}}\left(h_{c_i'},h^*_{\mathbb{P}^1_{a(i,j)}}\right)\\&+disc^\star_{\mathcal{L}}\left(\mathbb{P}^0_{a(i,j)},\mathbb{P}^1_{a(i,j)}\right)+\varepsilon\left(\mathbb{P}^0_{a(i,j)},\mathbb{P}^1_{a(i,j)}\right)\end{aligned}\tag{36}$$

Similarly, we have the following bounds :

$$\begin{aligned}&\mathcal{E}_{\mathbb{P}^1_{a(i,j)}}\left(h_{c_i'},h^*_{\mathbb{P}^1_{a(i,j)}}\right)\leq\mathcal{E}_{\mathbb{P}^2_{a(i,j)}}\left(h_{c_i'},h^*_{\mathbb{P}^2_{a(i,j)}}\right)\\&+disc^\star_{\mathcal{L}}\left(\mathbb{P}^1_{a(i,j)},\mathbb{P}^2_{a(i,j)}\right)+\varepsilon\left(\mathbb{P}^1_{a(i,j)},\mathbb{P}^2_{a(i,j)}\right)\end{aligned}\tag{37}$$

$\dots$

$$\begin{aligned}&\mathcal{E}_{\mathbb{P}^{c(i,j)-2}_{a(i,j)}}\left(h_{c_i'},f_{\mathbb{P}^{c(i,j)-2}_{a(i,j)}}\right)\leq\mathcal{E}_{\mathbb{P}^{c(i,j)-1}_{a(i,j)}}\left(h_{c_i'},h^*_{\mathbb{P}^{c(i,j)-1}_{a(i,j)}}\right)\\&+disc^\star_{\mathcal{L}}\left(\mathbb{P}^{c(i,j)-2}_{a(i,j)},\mathbb{P}^{c(i,j)-1}_{a(i,j)}\right)+\varepsilon\left(\mathbb{P}^{c(i,j)-2}_{a(i,j)},\mathbb{P}^{c(i,j)-1}_{a(i,j)}\right)\end{aligned}\tag{38}$$

$$\begin{aligned}&\mathcal{E}_{\mathbb{P}^{c(i,j)-1}_{a(i,j)}}\left(h_{c_i'},f_{\mathbb{P}^{c(i,j)-1}_{a(i,j)}}\right)\leq\mathcal{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}}\left(h_{c_i'},h^*_{\mathbb{P}^{c(i,j)}_{a(i,j)}}\right)\\&+disc^\star_{\mathcal{L}}\left(\mathbb{P}^{c(i,j)-1}_{a(i,j)},\mathbb{P}^{c(i,j)}_{a(i,j)}\right)+\varepsilon\left(\mathbb{P}^{c(i,j)-1}_{a(i,j)},\mathbb{P}^{c(i,j)}_{a(i,j)}\right)\end{aligned}\tag{39}$$

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

$$\begin{aligned}\mathcal{E}_{\mathcal{P}_{a(i,j)}}\left(h_{c_i'},f_{\mathcal{P}_{a(i,j)}}\right)\leq&\,\mathcal{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}}\left(h_{c_i'},h^*_{\mathbb{P}^{c(i,j)}_{a(i,j)}}\right)\\&+\sum_{k=-1}^{c(i,j)-1}\left\{disc^\star_{\mathcal{L}}\left(\mathbb{P}^k_{a(i,j)},\mathbb{P}^{k+1}_{a(i,j)}\right)\right.\\&\left.+\varepsilon\left(\mathbb{P}^k_{a(i,j)},\mathbb{P}^{k+1}_{a(i,j)}\right)\right\}8i\end{aligned}\tag{40}$$

where we also use $\mathbb{P}^{-1}_{a(i,j)}$ represent $\mathcal{P}_{a(i,j)}$. RHS of Eq. (40) is an upper bound to the target risk for a single task $\mathcal{P}_{a(i,j)}$ modelled by using the c_i'-th component. In the following, we consider all components C' that are trained more than once :

$$\begin{aligned}\sum_{i=1}^{|C'|}\sum_{j=1}^{\tilde{a}_i}\mathcal{E}_{\mathcal{P}_{a(i,j)}}\left(h_{c_i'},f_{\mathcal{P}_{a(i,j)}}\right)\leq&\sum_{i=1}^{|C'|}\sum_{j=1}^{\tilde{a}_i}\left\{\mathcal{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}}\left(h_{c_i'},h^*_{\mathbb{P}^{c(i,j)}_{a(i,j)}}\right)\right.\\&\left.+\sum_{k=-1}^{c(i,j)-1}\left(disc^\star_{\mathcal{L}}\left(\mathbb{P}^k_{a(i,j)},\mathbb{P}^{k+1}_{a(i,j)}\right)+\varepsilon\left(\mathbb{P}^k_{a(i,j)},\mathbb{P}^{k+1}_{a(i,j)}\right)\right)\right\}\end{aligned}\tag{41}$$

RHS of Eq. (41) is still an upper bound to the target risk of tasks modelled by the components that trained more than once. Therefore, $\mathcal{E}_{R'}$ in the paper, can be expressed by RHS of Eq. (41), which proves Theorem 3.

We provide additional analysis of the results of Theorem 3 in the following. Firstly, we rewrite Eq. (11) from the paper as :

$$\begin{aligned}&\frac{1}{t}\sum\nolimits_{i=1}^{|C|}\left\{\mathcal{E}_{\mathcal{P}_{a_i}}\left(h_{c_i},f_{\mathcal{P}_{a_i}}\right)\right\}\\&+\frac{1}{t}\sum\nolimits_{i=1}^{|C'|}\sum\nolimits_{j=1}^{\tilde{a}_i}\left\{\mathcal{R}_{\mathcal{P}_{a(i,j)}}\left(h_{c_i'},f_{\mathcal{P}_{a(i,j)}}\right)\right\}\leq\\&\sum\nolimits_{i=1}^{|C|}\left\{\mathcal{R}_{\tilde{\mathcal{P}}_{a_i}}\left(h_{c_i},h^*_{\tilde{\mathcal{P}}_{a_i}}\right)+\mathcal{E}_R\left(\mathcal{P}_{a_i},\tilde{\mathcal{P}}_{a_i}\right)\right\}\\&+\sum_{i=1}^{|C'|}\sum_{j=1}^{\tilde{a}_i}\left\{\mathcal{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}}\left(h_{c_i'},h^*_{\mathbb{P}^{c(i,j)}_{a(i,j)}}\right)\right.\\&\left.+\sum_{k=-1}^{c(i,j)-1}\left(\mathcal{R}_A\left(\mathbb{P}^k_{a(i,j)},\mathbb{P}^{k+1}_{a(i,j)}\right)\right)\right\}.\end{aligned}\tag{42}$$

We consider an extreme case where $\mathbf{M}$ only has a single component after LLL, ($|C'|=1$ and $|C|=0$). Then the first term in RHS of Eq. (42) disappears, resulting in :

$$\begin{aligned}&\frac{1}{t}\sum\nolimits_{i=1}^{|C'|}\sum\nolimits_{j=1}^{\tilde{a}_i}\mathcal{R}_{\mathcal{P}_{a(i,j)}}\left(h_{c_i'},f_{\mathcal{P}_{a(i,j)}}\right)\leq\\&\sum_{i=1}^{|C'|}\sum_{j=1}^{\tilde{a}_i}\left\{\mathcal{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}}\left(h_{c_i'},h^*_{\mathbb{P}^{c(i,j)}_{a(i,j)}}\right)\right.\\&\left.+\sum_{k=-1}^{c(i,j)-1}\left\{\mathcal{R}_A\left(\mathbb{P}^k_{a(i,j)},\mathbb{P}^{k+1}_{a(i,j)}\right)\right\}\right\}.\end{aligned}\tag{43}$$

In this case, when learning earlier tasks ($a(i,j)$ is small) tends to increase the number of error terms more than when learning more recent tasks ($a(i,j)$ is large), This is caused by the number of accumulated error terms $\mathcal{E}_A(\cdot)$ controlled by the number of times GR processes $c(i,j)=t-a(i,j)$ are used. In the opposite case where the number of components (K) is equal to the number of tasks (t), $|C'|=0$ and the mixture model has not accumulated errors. The GB for this case is :

$$\begin{aligned}\frac{1}{t}\sum\nolimits_{i=1}^{|C|}\mathcal{E}_{\mathcal{P}_{a_i}}\left(h_{c_i},f_{\mathcal{P}_{a_i}}\right)\leq&\sum\nolimits_{i=1}^{|C|}\left\{\mathcal{R}_{\tilde{\mathcal{P}}_{a_i}}\left(h_{c_i},h^*_{\tilde{\mathcal{P}}_{a_i}}\right)\right.\\&\left.+\mathcal{E}_R\left(\mathcal{P}_{a_i},\tilde{\mathcal{P}}_{a_i}\right)\right\}.\end{aligned}\tag{44}$$

where $|C|=K=t$. Then the lifelong learning problem is transformed to be the generalization problem under the generative modelling. This motivates us to propose a novel

dynamic mixture model which would not accumulate errors during LLL.

APPENDIX E
THE PROOF OF LEMMA 2

Similarly to the proof for Theorem 3, we firstly consider the components that are trained only once :

$$\mathcal{E}_C = \sum_{i=1}^{|C|} \Big\{ \mathcal{E}_{\tilde{\mathcal{P}}_{a_i}} \Big(h_{c_i}, h^*_{\tilde{\mathcal{P}}_{a_i}} \Big) + \mathcal{E}_R \Big(\mathcal{P}_{a_i}, \tilde{\mathcal{P}}_{a_i} \Big) \Big\}, \tag{45}$$

Then we add the KL divergence term and D_{diff} term in Eq. (45), resulting in :

$$\begin{aligned} \mathcal{E}_C = \sum_{i=1}^{|C|} \Big\{ & \mathcal{E}_{\tilde{\mathcal{P}}_{a_i}} \Big(h_{c_i}, h^*_{\tilde{\mathcal{P}}_{a_i}} \Big) + \mathbb{E}_{\tilde{\mathcal{P}}_{a_i}} KL \Big(p_{c_i}(\mathbf{z} \mid \mathbf{x}^S_{a_i}) \mid\mid p(\mathbf{z}) \Big) \\ & + D_{diff} \Big(\mathbf{x}^T_{a_i}, \mathbf{x}^S_{a_i} \Big) + \mathcal{E}_R \Big(\mathcal{P}_{a_i}, \tilde{\mathcal{P}}_{a_i} \Big) \Big\}, \end{aligned} \tag{46}$$

where $p_{c_i}(\mathbf{z} \mid \cdot)$ represents the variational distribution modelled by the inference model of the c_i-th component. $D_{diff}(\cdot, \cdot)$ is defined as :

$$\begin{aligned} D_{diff}(\mathbf{x}^T_{a_i}, \mathbf{x}^S_{a_i}) = \Big| & \mathbb{E}_{\mathcal{P}_{a_i}} KL(q_{\omega_{c_i}}(\mathbf{z} \mid \mathbf{x}^T_i) \mid\mid p(\mathbf{z})) \\ & - \mathbb{E}_{\tilde{\mathcal{P}}_{a_i}} KL(q_{\omega_{c_i}}(\mathbf{z} \mid \mathbf{x}^S_i) \mid\mid p(\mathbf{z})) \Big| \end{aligned} \tag{47}$$

We can rewrite Eq. (46) as the negative ELBO form :

$$\begin{aligned} \mathcal{E}_C = \sum_{i=1}^{|C|} \Big\{ & \mathbb{E}_{\tilde{\mathcal{P}}_i} \{ -\mathcal{L}_{ELBO} \Big(\mathbf{x}^S_{a_i}; h_{c_i} \Big) \} \\ & + D_{diff} \Big(\mathbf{x}^T_{a_i}, \mathbf{x}^S_{a_i} \Big) + \mathcal{E}_R \Big(\mathcal{P}_{a_i}, \tilde{\mathcal{P}}_{a_i} \Big) \Big\}, \end{aligned} \tag{48}$$

where $\mathcal{L}_{ELBO}(\cdot; h_{c_i})$ represents the ELBO estimated by the c_i-th component. Secondly, we consider the components that are trained more than once :

$$\begin{aligned} \mathcal{E}_{R'} = \sum_{i=1}^{|C'|} \sum_{j=1}^{\tilde{a}_i} \{ & \mathcal{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}} \Big(h_{c'_i}, f_{\mathbb{P}^{c(i,j)}_{a(i,j)}} \Big) \\ & + \mathcal{E}_R \Big(\mathcal{P}_{a(i,j)}, \mathbb{P}^{c(i,j)}_{a(i,j)} \Big) \}, \end{aligned} \tag{49}$$

We then add the KL divergence term and D_{diff} term in eq. (49), resulting in :

$$\begin{aligned} \mathcal{E}_{R'} = \sum_{i=1}^{|C'|} \sum_{j=1}^{\tilde{a}_i} \Big\{ & \mathcal{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}} \Big(h_{c'_i}, f_{\mathbb{P}^{c(i,j)}_{a(i,j)}} \Big) \\ & + \mathbb{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}} KL \Big(p_{c'_i}(\mathbf{z} \mid \mathbf{x}^t_{a(i,j)}) \mid\mid p(\mathbf{z}) \Big) \\ & + D_{diff} \Big(\mathbf{x}^T_{a(i,j)}, \mathbf{x}^t_{a(i,j)} \Big) + \mathcal{E}_R \Big(\mathcal{P}_{a(i,j)}, \mathbb{P}^{c(i,j)}_{a(i,j)} \Big) \Big\}, \end{aligned} \tag{50}$$

where the conditional distribution $p_{c'_i}(\mathbf{z}|\mathbf{x}^t_{a(i,j)})$ is modelled by the inference model of the c'_i-th component. Then we rewrite Eq. (50) as the negative ELBO form :

$$\begin{aligned} \mathcal{E}_{R'} = \sum_{i=1}^{|C'|} \sum_{j=1}^{\tilde{a}_i} & \mathbb{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}} \Big\{ -\mathcal{L}_{ELBO} \Big(\mathbf{x}^t_{a(i,j)}; h_{c'_i} \Big) \\ & + D_{diff} \Big(\mathbf{x}^T_{a(i,j)}, \mathbf{x}^t_{a(i,j)} \Big) + \mathcal{E}_R \Big(\mathcal{P}_{a(i,j)}, \mathbb{P}^{c(i,j)}_{a(i,j)} \Big) \Big\}, \end{aligned} \tag{51}$$

We summarize all KL divergence and D_{diff} terms :

$$\begin{aligned} & \sum_{i=1}^{|C|} \Big\{ \mathbb{E}_{\tilde{\mathcal{P}}_{a_i}} KL \Big(p_{c_i}(\mathbf{z} \mid \mathbf{x}^S_{a_i}) \mid\mid p(\mathbf{z}) \Big) + D_{diff} \Big(\mathbf{x}^T_{a_i}, \mathbf{x}^S_{a_i} \Big) \Big\} \\ & + \sum_{i=1}^{|C'|} \sum_{j=1}^{\tilde{a}_i} \Big\{ \mathbb{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}} KL \Big(p_{c'_i}(\mathbf{z} \mid \mathbf{x}^t_{a(i,j)}) \mid\mid p(\mathbf{z}) \Big) \\ & + D_{diff} \Big(\mathbf{x}^T_{a(i,j)}, \mathbf{x}^t_{a(i,j)} \Big) \Big\}, \end{aligned} \tag{52}$$

We also know that :

$$\begin{aligned} & \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}_{\mathcal{P}_i} KL \Big(p(\mathbf{z} \mid \mathbf{x}^T_i) \mid\mid p(\mathbf{z}) \Big) \leq \\ & \frac{1}{t} \sum_{i=1}^{|C|} \Big\{ \mathbb{E}_{\tilde{\mathcal{P}}_{a_i}} KL \Big(p(\mathbf{z} \mid \mathbf{x}^S_{a_i}) \mid\mid p(\mathbf{z}) \Big) + D_{diff} \Big(\mathbf{x}^T_{a_i}, \mathbf{x}^S_{a_i} \Big) \Big\} \\ & + \frac{1}{t} \sum_{i=1}^{|C'|} \sum_{j=1}^{\tilde{a}_i} \Big\{ \mathbb{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}} KL \Big(p(\mathbf{z} \mid \mathbf{x}^t_{a(i,j)}) \mid\mid p(\mathbf{z}) \Big) \\ & + D_{diff} \Big(\mathbf{x}^T_{a(i,j)}, \mathbf{x}^t_{a(i,j)} \Big) \Big\} \end{aligned} \tag{53}$$

where we omit the subscript for $p(\mathbf{z} \mid \cdot)$ for simplicity. We then consider the inequality from Eq. (53) into Eq. (11) of the paper and we have :

$$\begin{aligned} & \frac{1}{t} \sum_{i=1}^{|C|} \mathbb{E}_{\mathcal{P}_{a_i}} \Big\{ -\mathcal{L}_{ELBO} \Big(\mathbf{x}^T_{a_i}; h_{c_i} \Big) \Big\} \\ & + \frac{1}{t} \sum_{i=1}^{|C'|} \sum_{j=1}^{\tilde{a}_i} \mathbb{E}_{\mathcal{P}_{a(i,j)}} \Big\{ -\mathcal{L}_{ELBO} \Big(\mathbf{x}^T_{a(i,j)}; h_{c'_i} \Big) \Big\} \leq \\ & \frac{1}{t} \sum_{i=1}^{|C|} \Big\{ \mathbb{E}_{\tilde{\mathcal{P}}_{a_i}} \Big\{ -\mathcal{L}_{ELBO} \Big(\mathbf{x}^S_{a_i}; h_{c_i} \Big) \Big\} \Big\} \\ & + \frac{1}{t} \sum_{i=1}^{|C'|} \sum_{j=1}^{|A'_{c'_i}|} \Big\{ \mathbb{E}_{\mathbb{P}^{c(i,j)}_{a(i,j)}} \Big\{ -\mathcal{L}_{ELBO} \Big(\mathbf{x}^t_{a(i,j)}; h_{c'_i} \Big) \Big\} \Big\} \\ & + \frac{1}{t} \{ \mathcal{R}^{II}_{R'} + \mathcal{R}^{II}_C + D_{diff}* \} \end{aligned} \tag{54}$$

This proves Lemma 2. We should also observe that $D_{diff}*$ is expressed by :

$$\begin{aligned} D_{diff}* = & \sum_{i=1}^{|C|} \Big\{ D_{diff} \Big(\mathbf{x}^T_{a_i}, \mathbf{x}^S_{a_i} \Big) \Big\} \\ & + \sum_{i=1}^{|C'|} \sum_{j=1}^{\tilde{a}_i} \Big\{ D_{diff} \Big(\mathbf{x}^T_{a(i,j)}, \mathbf{x}^t_{a(i,j)} \Big) \Big\}, \end{aligned} \tag{55}$$

$\mathcal{R}^{II}_{R'}$ and $\mathcal{R}^{II}_C$ are defined as :

$$\mathcal{R}^{II}_{R'} = \sum\nolimits_{i=1}^{|C|} \Big\{ \mathcal{E}_R \Big(\mathcal{P}_{a_i}, \tilde{\mathcal{P}}_{a_i} \Big) \Big\} \tag{56}$$

$$\mathcal{R}^{II}_C = \sum\nolimits_{i=1}^{|C'|} \Big\{ \sum\nolimits_{j=1}^{\tilde{a}_i} \Big\{ \mathcal{E}_R \Big(\mathcal{P}_{a(i,j)}, \mathbb{P}^{c(i,j)}_{a(i,j)} \Big) \Big\} \Big\} \tag{57}$$

Based on the above results, in the following, we derive the risk bound of the mixture model to NLL :

$$\begin{aligned}
&\frac{1}{t}\sum_{i=1}^{|C|}\mathbb{E}_{\mathcal{P}_{a_i}}\Big\{-\log p_{c_i}\left(\mathbf{x}_{a_i}^T\right)\Big\}\\
&+\frac{1}{t}\sum_{i=1}^{|C'|}\sum_{j=1}^{\tilde{a}_i}\Big\{-\log p_{c'_i}\left(\mathbf{x}_{a(i,j)}^T\right)\Big\}\leq\\
&\frac{1}{t}\sum_{i=1}^{|C|}\Big\{\mathbb{E}_{\tilde{\mathcal{P}}_{a_i}}\Big\{-\mathcal{L}_{ELBO}\left(\mathbf{x}_{a_i}^S;h_{c_i}\right)\Big\}\Big\}\\
&+\frac{1}{t}\sum_{i=1}^{|C'|}\sum_{j=1}^{|A'_{c'_i}|}\Big\{\mathbb{E}_{\mathbb{P}_{a(i,j)}^{c(i,j)}}\Big\{-\mathcal{L}_{ELBO}\left(\mathbf{x}_{a(i,j)}^t;h_{c'_i}\right)\Big\}\Big\}\\
&+\frac{1}{t}\Big\{\mathcal{R}_{R'}^{II}+\mathcal{R}_{C}^{II}+D_{diff}*\Big\}
\end{aligned}\tag{58}$$

where $\log p_{c_i}(\cdot)$ represents the sample log-likelihood (model likelihood) under the c_i-th component.

APPENDIX F
PROOF OF PROPOSITION 2

In order to measure the forgetting behaviour of a GAN for each task learning, we need to define the individual approximation distribution related to each task. Let us define the approximation distribution $\mathbb{P}_i^j$ formed by the sampling process $\mathbf{x}\sim\mathbb{P}^j$ if $I_{\mathcal{T}}(\mathbf{x})=i$. $\mathbb{P}_i^j$ is the probabilistic representation of the generated data related to the i-th task where j represents that the GAN model has been trained on a number j of tasks. we use $\mathbb{P}_i^{(i-1)}$ to represent $\tilde{\mathcal{P}}_i$ for simplicity. Therefore, for the i-th task, we have the following bound :

$$\mathcal{E}_{\mathbb{P}_i^{(i-1)}}\left(h,f_{\mathcal{P}_i}\right)\leq\mathcal{E}_{\mathbb{P}_i^{i}}\left(h,f_{\mathcal{P}_i}\right)+\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_i^{i},\mathbb{P}_i^{(i-1)}\right),\tag{59}$$

and

$$\mathcal{E}_{\mathbb{P}_i^{i}}\left(h,f_{\mathcal{P}_i}\right)\leq\mathcal{E}_{\mathbb{P}_i^{(i+1)}}\left(h,f_{\mathcal{P}_i}\right)+\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_i^{(i+1)},\mathbb{P}_i^{i}\right),\tag{60}$$

In the following, we treat $\mathbb{P}_i^{(i+1)}$ as the target distribution and $\mathbb{P}_i^{(i+2)}$ as the source distribution, we have :

$$\begin{aligned}
\mathcal{E}_{\mathbb{P}_i^{(i+1)}}\left(h,f_{\mathcal{P}_i}\right)\leq&\,\mathcal{E}_{\mathbb{P}_i^{(i+2)}}\left(h,f_{\mathcal{P}_i}\right)\\
&+\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_i^{(i+2)},\mathbb{P}_i^{(i+1)}\right),
\end{aligned}\tag{61}$$

We repeat this process, and through induction :

$$\begin{aligned}
&\mathcal{E}_{\mathbb{P}_i^{((i+2)}}\left(h,f_{\mathcal{P}_i}\right)\leq\mathcal{E}_{\mathbb{P}_i^{((i+3)}}\left(h,f_{\mathcal{P}_i}\right)+\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_i^{(i+3)},\mathbb{P}_i^{(i+2)}\right)\\
&\qquad\cdots\\
&\qquad\cdots\\
&\mathcal{E}_{\mathbb{P}_i^{t-1}}\left(h,f_{\mathcal{P}_i}\right)\leq\mathcal{E}_{\mathbb{P}_i^{t}}\left(h,f_{\mathcal{P}_i}\right)+\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_i^{t},\mathbb{P}_i^{t-1}\right)
\end{aligned}\tag{62}$$

We then sum up all inequalities, resulting in :

$$\begin{aligned}
\mathcal{E}_{\mathbb{P}_i^{(i-1)}}\left(h,f_{\mathcal{P}_i}\right)\leq&\,\mathcal{E}_{\mathbb{P}_i^{t}}\left(h,f_{\mathcal{P}_i}\right)\\
&+\sum_{j=i}^{t}\Big\{\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_i^{j-1},\mathbb{P}_i^{j}\right)\Big\},
\end{aligned}\tag{63}$$

We can observe that the left hand side (LHS) of Eq. (63) is also an upper bound to the target risk of the model at the i-th task :

$$\mathcal{E}_{\mathcal{P}_i}\left(h,f_{\mathcal{P}_i}\right)\leq\mathcal{E}_{\mathbb{P}_i^{(i-1)}}\left(h,f_{\mathcal{P}_i}\right)+\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_i^{(i-1)},\mathcal{P}_i\right),\tag{64}$$

By comparing Eq. (64) and Eq. (63), we have a GB for the i-th task :

$$\begin{aligned}
\mathcal{E}_{\mathcal{P}_i}\left(h,f_{\mathcal{P}_i}\right)\leq&\,\mathcal{E}_{\mathbb{P}_i^{t}}\left(h,f_{\mathcal{P}_i}\right)+\sum_{j=i}^{t}\Big\{\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_i^{j-1},\mathbb{P}_i^{j}\right)\Big\}\\
&+\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_i^{i-1},\mathcal{P}_i\right),
\end{aligned}\tag{65}$$

Then we can easily obtain a GB for all tasks based on Eq. (65).

$$\begin{aligned}
\sum_{k=1}^{t}\mathcal{E}_{\mathcal{P}_k}\left(h,f_{\mathcal{P}_k}\right)\leq&\sum_{k=1}^{t}\Big\{\mathcal{E}_{\mathbb{P}_k^{t}}\left(h,f_{\mathcal{P}_k}\right)\\
&+\sum_{j=k}^{t}\left\{\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_k^{j-1},\mathbb{P}_k^{j}\right)\right\}\\
&+\mathcal{L}_{\text{disc}}^{\star}\left(\mathbb{P}_k^{k-1},\mathcal{P}_i\right)\Big\},
\end{aligned}\tag{66}$$

This proves Proposition 2.

APPENDIX G
IMPORTANCE SAMPLING

The main idea of Importance Weighted Autoencoder (IWELBO) [4] is to allow the recognition network to generate multiple samples during the optimization leading to a better modelling of the posterior probabilities. The corresponding ELBO for sampling K' samples is defined as :

$$\mathcal{L}_{ELBO_{K'}}\left(\mathbf{x};\mathcal{M}\right)=\mathbb{E}_{\mathbf{z}_1,..,\mathbf{z}_{K'}\sim q(\mathbf{z}|\mathbf{x})}\left[\log\frac{1}{K'}\sum_{i=1}^{K'}\frac{p\left(\mathbf{x},\mathbf{z}_i\right)}{q\left(\mathbf{z}_i|\mathbf{x}\right)}\right]\tag{67}$$

where K' is the number of weighted samples and $K'=1$ is equivalent to the standard ELBO. In order to calculate $w_i=p(\mathbf{x},\mathbf{z}_i)/q(\mathbf{z}_i|\mathbf{x})$ in practice, we rewrite w_i as $\exp(\log w_i)$. By calculating the right hand side of Eq. (67) requires to estimate each individual $\mathbb{E}_{z_i}\log w_i$ which is a standard ELBO.

In the following, we extend this IWELBO to the LLL setting. From Lemma 1, we know that :

$$\begin{aligned}
&\frac{1}{t}\sum\nolimits_{i=1}^{t}\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\Big[-\mathcal{L}_{ELBO}\left(\mathbf{x}_i^T;h\right)\Big]\leq\\
&\mathbb{E}_{\mathbf{x}^t\sim\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\Big[-\mathcal{L}_{ELBO}\left(\tilde{\mathbf{x}}^t;h\right)\Big]\\
&+|KL_1-KL_2|+\text{Err}^a+\text{Err}^d.
\end{aligned}\tag{68}$$

where $\mathcal{L}_{ELBO}(\cdot)$ has the form according to [5]. We can rewrite the above equation, by considering importance sampling, as :

$$\begin{aligned}
&\frac{1}{t}\sum\nolimits_{i=1}^{t}\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\Big[-\log p\left(\mathbf{x}_i^T\right)\Big]\leq\\
&\mathbb{E}_{\mathbf{x}^t\sim\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\Big[-\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}\left[\log\frac{p\left(\tilde{\mathbf{x}}^t,\mathbf{z}\right)}{q\left(\mathbf{z}\mid\mathbf{x}\right)}\right]\Big]\\
&+|KL_1-KL_2|+\text{Err}^a+\text{Err}^d.
\end{aligned}\tag{69}$$

According to $-\log p(\mathbf{x}) \leq -\mathcal{L}_{ELBO_{K'+1}}(\mathbf{x};\mathcal{M}) \leq -\mathcal{L}_{ELBO_{K'}}(\mathbf{x};\mathcal{M})$ [4], we have $-\mathcal{L}_{ELBO_{K'+1}}(\tilde{\mathbf{x}}^t;\mathcal{M}) \leq -\mathcal{L}_{ELBO_{K'}}(\tilde{\mathbf{x}}^t;\mathcal{M})$, based on the assumation that $\mathbb{P}^{t-1}$ is fixed. We note that $\mathcal{M}$ represents the model and h is the hypothesis of $\mathcal{M}$. We assume that when the $h^\star$ is an optimal solution for $-\log p(\tilde{\mathbf{x}}^t)$ and we have $-\log p(\tilde{\mathbf{x}}^t) \leq -\mathcal{L}_{ELBO}(\tilde{\mathbf{x}}^t;h^\star) \leq -\mathcal{L}_{ELBO_{K'}}(\tilde{\mathbf{x}}^t;h)$. In the following, we rewrite Eq. (68) by using $h^\star$:

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\Big[-\mathcal{L}_{ELBO}\left(\mathbf{x}_i^T;h^\star\right)\Big] \leq \\ &\mathbb{E}_{\mathbf{x}^t\sim\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\Big[-\mathcal{L}_{ELBO}\left(\tilde{\mathbf{x}}^t;h^\star\right)\Big] \\ &+|KL_1-KL_2|+\mathrm{Err}^a+\mathrm{Err}^d.\end{aligned} \tag{70}$$

We observe that $|KL_1 - KL_2|$ of Eq. (70) is estimated by $h^\star$ ($\{\theta_\star, \omega_\star\}$ are the corresponding model parameters). We can replace the first term in RHS of Eq. (70) by using $\mathcal{L}_{ELBO_{K'}}(\tilde{\mathbf{x}}^t;h)$, resulting in :

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\Big[-\mathcal{L}_{ELBO}\left(\mathbf{x}_i^T;h^\star\right)\Big] \leq \\ &\mathbb{E}_{\mathbf{x}^t\sim\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\Big[-\mathcal{L}_{ELBO_{K'}}(\tilde{\mathbf{x}}^t;h)\Big] \\ &+|KL_1-KL_2|+\mathrm{Err}^a+\mathrm{Err}^d.\end{aligned} \tag{71}$$

LHS of Eq. (71) is an upper bound to $\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}[-\log p_{\theta^\star}(\mathbf{x}_i^T)]$ and we rewrite Eq. (71) as :

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\Big[-\log p_{\theta^\star}(\mathbf{x}_i^T)\Big] \leq \\ &\mathbb{E}_{\mathbf{x}^t\sim\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\Big[-\mathcal{L}_{ELBO_{K'}}(\tilde{\mathbf{x}}^t;h)\Big] \\ &+|KL_1-KL_2|+\mathrm{Err}^a+\mathrm{Err}^d.\end{aligned} \tag{72}$$

We then decompose the first term in RHS of Eq. (72), and we have :

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\Big[-\log p_{\theta^\star}\left(\mathbf{x}_i^T\right)\Big] \leq \\ &\mathbb{E}_{\mathbf{x}^t\sim\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left[-\mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_{K'}\sim q(\mathbf{z}|\mathbf{x})}\left[\log\frac{1}{K'}\sum_{i=1}^{K'}\frac{p_\theta\left(\tilde{\mathbf{x}}^t,\mathbf{z}_i\right)}{q_\omega\left(\mathbf{z}_i\mid\mathbf{x}\right)}\right]\right] \\ &+|KL_1-KL_2|+\mathrm{Err}^a+\mathrm{Err}^d.\end{aligned} \tag{73}$$

If we do not consider Err^a and Err^d as in Lemma 1 (See details in Appendix C), Eq. (73) can be :

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_{\mathbf{x}_i^T\sim\mathcal{P}_i}\left[-\log p\left(\mathbf{x}_i^T\right)\right] \leq \\ &\mathbb{E}_{\tilde{\mathbf{x}}^t\sim\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left[-\mathbb{E}_{\mathbf{z}_1\ldots,\mathbf{z}_{K'}\sim q(\mathbf{z}|\mathbf{x})}\left[\log\frac{1}{K'}\sum_{i=1}^{K'}\frac{p\left(\tilde{\mathbf{x}}^t,\mathbf{z}_i\right)}{q\left(\mathbf{z}_i\mid\mathbf{x}\right)}\right]\right] \\ &+|KL_1-KL_2|+\mathcal{R}_A\left(\mathcal{P}_{(1:t)},\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t\right).\end{aligned} \tag{74}$$

where we omit the subscript (the model's parameters) for Eq. (74) for the sake of simplification. It notes that $h \in \mathcal{H}$ is the model and its parameters are $\{\theta, \omega\}$ optimized by Eq.(73). We call RHS of Eq. (73) as $\mathcal{L}_{LELBO_{K'}}$ and when $K' = 1$, $\mathcal{L}_{LELBO_{K'}}$ is equal to $\mathcal{L}_{LELBO}$ (RHS of Eq. (69)). Based on the assumption that $\mathbb{P}^{t-1} \otimes \tilde{\mathcal{P}}_t$ is fixed, we have $\mathcal{L}_{LELBO_{K'+1}} \leq \mathcal{L}_{LELBO_{K'}}$. We can observe that the tightness of ELBO on the marginal log-likelihood of the source distribution $\mathbb{P}^{t-1} \otimes \tilde{\mathcal{P}}_t$ can not guarantee a tight GB on the marginal log-likelihood of the target distribution. However, the tightness of GB is largely depending on the discrepancy distance between the evolved source and target distribution.

APPENDIX H
DERIVATION OF $\mathcal{L}_{MELBO}$ (THEOREM 4)

As illustrated in Fig.2 from the paper, where we show the graph structure implementing DEGM, we have a number K of basic nodes in DEGM. When building a new specific node for learning a new task, we derive the main objective function showing as follows.

Theorem 4. When a new specific node ($(t+1)$-th node) is built for learning the $(t+1)$-th task. This specific node connected with all basic nodes can be seen as a sub-graph model which can be trained by a valid lower bound (ELBO).

Proof. We start by considering the KL divergence [6] :

$$KL\left[Q\left(\mathbf{z}\right)||p\left(\mathbf{z}|\mathbf{x}\right)\right] = \mathbb{E}_{z\sim Q(\mathbf{z})}\left[\log Q\left(\mathbf{z}\right) - \log p\left(\mathbf{z}|\mathbf{x}\right)\right], \tag{75}$$

where $Q(\mathbf{z})$ is the variational distribution. We can rewrite the above equation as :

$$\begin{aligned}KL\left[Q\left(\mathbf{z}\right)||p\left(\mathbf{z}|\mathbf{x}\right)\right] = \mathbb{E}_{z\sim Q(\mathbf{z})}[&\log Q\left(\mathbf{z}\right) \\ &- \log p\left(\mathbf{x}|\mathbf{z}\right) - \log p\left(z\right)] \\ &+ \log p\left(\mathbf{x}\right)\end{aligned} \tag{76}$$

And we have :

$$\begin{aligned}\log p\left(\mathbf{x}\right) - KL\left[Q\left(\mathbf{z}\right)||p\left(\mathbf{z}|\mathbf{x}\right)\right] = \mathbb{E}_{z\sim Q(\mathbf{z})}\left[\log p\left(\mathbf{x}|\mathbf{z}\right)\right] \\ - KL\left[Q\left(\mathbf{z}\right)||p\left(\mathbf{z}\right)\right],\end{aligned} \tag{77}$$

where the right hand side is also called evidence lower bound (ELBO). We particularly focus on the KL term $KL\left[Q\left(\mathbf{z}\right)||p\left(\mathbf{z}\right)\right]$ which has the following from :

$$KL\left[Q\left(\mathbf{z}\right)||p\left(\mathbf{z}\right)\right] = \int q\left(\mathbf{z}\right)\log\frac{q\left(\mathbf{z}\right)}{p\left(\mathbf{z}\right)}d\mathbf{z}, \tag{78}$$

where $q(\mathbf{z})$ is the density of $Q(\mathbf{x})$. Since we have K components and we consider the $q(\mathbf{z})$ to be mixture density function $q\left(\mathbf{z}\right) = \sum_{i=1}^{w}\pi_i q_{\tilde{\omega}_{\mathcal{SG}(i)}} \circ q_{\omega'_{(t+1)}}\left(\mathbf{z}|\mathbf{x}\right)$ where π_i is the weight. We then rewrite Eq. (78) as :

$$\begin{aligned}&KL\left[Q\left(\mathbf{z}\right)||p\left(\mathbf{z}\right)\right] = \\ &\int\left(\sum_{i=1}^{K}\pi_i q_{\tilde{\omega}_{\mathcal{GI}(i)}} \circ q_{\omega'_{(t+1)}}\left(\mathbf{z}|\mathbf{x}\right)\right)\log\frac{q\left(\mathbf{z}\right)}{p\left(\mathbf{z}\right)}d\mathbf{z} \\ &= \sum_{i=1}^{K}\pi_i\int q_{\tilde{\omega}_{\mathcal{GI}(i)}} \circ q_{\omega'_{(t+1)}}\left(\mathbf{z}|\mathbf{x}\right)\log\frac{q\left(\mathbf{z}\right)}{p\left(\mathbf{z}\right)}d\mathbf{z}\end{aligned} \tag{79}$$

We then add the term $q_{\tilde{\omega}_{\mathcal{SG}(i)}} \circ q_{\omega'_{(t+1)}}\left(\mathbf{z}|\mathbf{x}\right)/q_{\tilde{\omega}_{\mathcal{SG}(i)}} \circ q_{\omega'_{(t+1)}}\left(\mathbf{z}|\mathbf{x}\right)$ to Eq. (79), resulting in :

$$\begin{aligned}&KL\left[Q\left(\mathbf{z}\right)||p\left(\mathbf{z}\right)\right] = \\ &\sum_{i=1}^{K}\pi_i\int q_{\tilde{\omega}_{\mathcal{GI}(i)}} \circ q_{\omega'_{(t+1)}}\left(\mathbf{z}|\mathbf{x}\right)\log\{\frac{q\left(\mathbf{z}\right)}{p\left(\mathbf{z}\right)} \\ &\times\frac{q_{\tilde{\omega}_{\mathcal{GI}(i)}} \circ q_{\omega'_{(t+1)}}\left(\mathbf{z}|\mathbf{x}\right)}{q_{\tilde{\omega}_{\mathcal{GI}(i)}} \circ q_{\omega'_{(t+1)}}\left(\mathbf{z}|\mathbf{x}\right)}\}d\mathbf{z}\end{aligned} \tag{80}$$

We rewrite the above equation as :

$$\sum_{i=1}^{K}\pi_i\int q_{\tilde{\omega}_{\mathcal{GI}(i)}}\circ q_{\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})\times\log\{\frac{q_{\tilde{\omega}_{\mathcal{GI}(i)}}\circ q_{\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\times\frac{q(\mathbf{z})}{q_{\tilde{\omega}_{\mathcal{GI}(i)}}\circ q_{\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})}\}d\mathbf{z}=\sum_{i=1}^{K}\pi_i\int q_{\tilde{\omega}_{\mathcal{GI}(i)}}\circ q_{\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})\times\left(\log\frac{q_{\tilde{\omega}_{\mathcal{GI}(i)}}\circ q_{\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}+\log\frac{q(\mathbf{z})}{q_{\tilde{\omega}_{\mathcal{GI}(i)}}\circ q_{\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})}\right)d\mathbf{z} \tag{81}$$

Then we can rewrite the above equation as KL terms :

$$KL\left[Q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right]=\sum_{i=1}^{K}\left(\pi_i KL\left[Q_{\tilde{\omega}_{\mathcal{GI}(i)}\circ\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]\right)-\sum_{i=1}^{K}\left(\pi_i KL\left[Q_{\tilde{\omega}_{\mathcal{GI}(i)}\circ\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]\right) \tag{82}$$

where $q_{\tilde{\omega}_{\mathcal{GI}(i)}}\circ q_{\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})$ is the density form of $Q_{\tilde{\omega}_{\mathcal{GI}(i)}\circ\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})$. We replace the expression of $KL\left[Q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right]$ into Eq. (77), resulting in :

$$\log p(\mathbf{x})-KL\left[Q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right]=\mathbb{E}_{z\sim Q(\mathbf{z})}\left[\log p(\mathbf{x}|\mathbf{z})\right]-\sum_{i=1}^{K}\left(\pi_i KL\left[Q_{\tilde{\omega}_{\mathcal{GI}(i)}\circ\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]\right)+\sum_{i=1}^{K}\left(\pi_i KL\left[Q_{\tilde{\omega}_{\mathcal{GI}(i)}\circ\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})||Q(\mathbf{z})\right]\right) \tag{83}$$

where we move the last term of the right hand side to the left hand side, resulting in :

$$\log p(\mathbf{x})-KL\left[Q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right]-\sum_{i=1}^{K}\left(\pi_i KL\left[Q_{\tilde{\omega}_{\mathcal{GI}(i)}\circ\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})||Q(\mathbf{z})\right]\right)=\mathbb{E}_{z\sim Q(\mathbf{z})}\left[\log p(\mathbf{x}|\mathbf{z})\right]-\sum_{i=1}^{K}\left(\pi_i KL\left[Q_{\tilde{\omega}_{\mathcal{GI}(i)}\circ\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]\right) \tag{84}$$

We know that KL terms are always larger or equal to zero and the right hand side of Eq. (84) is a lower bound to the sample log-likelihood. Finally, the objective function for training DEGM is to maximize this lower bound :

$$\mathbb{E}_{z\sim Q(\mathbf{z})}\left[\log p(\mathbf{x}|\mathbf{z})\right]-\sum_{i=1}^{K}\left(\pi_i KL\left[Q_{\tilde{\omega}_{\mathcal{GI}(i)}\circ\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]\right) \tag{85}$$

where in the first term, $\mathbf{z}$ is sampled from the mixture distribution $Q(\mathbf{z})=\sum_{i=1}^{K}\pi_i Q_{\tilde{\omega}_{\mathcal{GI}(i)}\circ\omega'_{(t+1)}}(\mathbf{z}|\mathbf{x})$. In order to reuse the parameters and transferable information from all basic nodes, we consider the following implementations for the decoder. When calculating the first term, we build an input layer on the top of the decoder. This input layer is used as the identity function and is connected with each sub-decoder $g_{\tilde{\theta}_{\mathcal{GI}(i)}}(\mathbf{z})$ of each basic node, represented as one layer or a module in a single decoder. Then we obtain the intermediate data representation $\tilde{\mathbf{x}}=\sum_{i=1}^{K}\pi_i g_{\tilde{\theta}_{\mathcal{GI}(i)}}(\mathbf{z})$ which is used as the input for the newly created sub-decoder $g_{\theta'_{(t+1)}}(\tilde{\mathbf{x}})$. Therefore, we treat the intermediate data representation $\tilde{\mathbf{x}}$ as the information between two layers in a single decoder and do not consider $\tilde{\mathbf{x}}$ to be the random variable. In this case we rewrite the decoding distribution $p(\mathbf{x}|\mathbf{z})$ as :

$$p(\mathbf{x}|\mathbf{z})=p_{\theta'_{(t+1)}\circ\{\tilde{\theta}_{\mathcal{GI}(1)},\ldots,\tilde{\theta}_{\mathcal{GI}(K)}\}}(\mathbf{x}\,|\,\mathbf{z})$$

. Therefore, we rewrite Eq. (85) as :

$$\mathcal{L}_{MELBO}\left(\mathbf{x};\mathcal{M}_{(t+1)}\right)=\mathbb{E}_{Q(\mathbf{z})}\left[\log p_{\theta'_{(t+1)}\circ\{\tilde{\theta}_{\mathcal{GI}(1)},\ldots,\tilde{\theta}_{\mathcal{GI}(K)}\}}(\mathbf{x}\mid\mathbf{z})\right]-\sum_{i=1}^{K}\left(\pi_i KL\left[q_{\tilde{\omega}_{\mathcal{GI}(i)}}\circ q_{\omega'_{(t+1)}}(\mathbf{z}\mid\mathbf{x})\;||\;p(\mathbf{z}_i)\right]\right) \tag{86}$$

Eq. (86) shows that we can implement existing variational inference mechanisms such as using a more expressive posterior [7], [8], [9], important sampling [4] and Semi-Implicit Variational Inference [10], [11] in our framework to enable for lifelong learning.

APPENDIX I
PERFORMANCE CRITERION

In unsupervised image reconstruction, we adapt the structural similarity index measure (SSIM) [12], the Mean Squared Error (MSE) and the Peak-Signal-to-Noise Ratio (PSNR) [12] in order to evaluate the image reconstruction quality. The calculation form of the MSE, SSIM and PSNR criteria are provided as follows:

$$\mathrm{MSE}(\mathbf{X}_{test},\mathbf{X}'_{test})=\frac{1}{n}\sum_{i}^{n}\{||\mathbf{x}_i-\mathbf{x}'_i||^2\}, \tag{87}$$

$$\mathrm{SSIM}(\mathbf{x},\mathbf{x}')=\frac{(2\mu_{\mathbf{x}}\mu_{\mathbf{x}'}+c_1)(2\sigma_{\mathbf{x}\mathbf{x}'}+c_2)}{(\mu_{\mathbf{x}}^2+\mu_{\mathbf{x}'}^2+c_1)(\sigma_{\mathbf{x}}^2+\sigma_{\mathbf{x}'}^2+c_2)}, \tag{88}$$

$$\mathrm{PSNR}(\mathbf{x},\mathbf{x}')=10\log_{10}\frac{\max(\mathbf{x})^2}{\mathrm{MSE}(\mathbf{x},\mathbf{x}')}, \tag{89}$$

where $\mathbf{x}$ and $\mathbf{x}'$ denote the real testing and reconstructed image, respectively. $\mathbf{X}_{test}$ and $\mathbf{X}'_{test}$ are the testing dataset and the reconstructed image dataset, respectively. $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{x}'}$ represent the pixel sample mean of $\mathbf{x}$ and $\mathbf{x}'$, respectively. $\sigma_{\mathbf{x}}^2$ and $\sigma_{\mathbf{x}'}^2$ are the variance of $\mathbf{x}$ and $\mathbf{x}'$, respectively. $\mathrm{MSE}(\cdot,\cdot)$ is the mean square error. $c_1=(k_1L_{\mathrm{image}})^2$ and $c_1=(k_2L_{\mathrm{image}})^2$ are two variables where L_{image} is the dynamic range of the pixel-values and $k_1=0.01$ and $k_2=0.03$. We employ the skimage library to implement the PSNR and SSIM criteria.

APPENDIX J
ANALYSIS FOR OTHER GENERATIVE MODELS

In the following we extend the GB defined above to other types of generative models.

J.1 Generative Adversarial Nets (GANs)

The discrepancy distance was used in GANs [13] for enabling the generation of realistic data by matching the generator's distribution to the real data distribution in the discriminator module. However, the discrepancy distance for GANs has not been studied in the context of lifelong learning. In this section, we derive a risk bound for the GAN model and provide the analysis for the forgetting behaviour during lifelong learning. Following from [13], we define $\mathcal{L}_{\mathcal{H}}=\{\mathcal{L}(h(\mathbf{x}),h'(\mathbf{x})) : h,h' \in \mathcal{H}\}$ as a family of discriminators which is used in the estimation of the discrepancy distance. Adlam *et al.* [13] derived a risk bound for the GAN model when it is trained on a single target distribution :

$$\mathcal{E}_{\mathcal{P}_i}\left(h, f_{\mathcal{P}_i}\right) \leq \mathcal{E}_{\mathbb{P}^1}\left(h, f_{\mathcal{P}_i}\right) + \mathcal{L}^{\star}_{\text{disc}}\left(\mathcal{P}_1, \mathbb{P}^1\right) , \tag{90}$$

where $f_{\mathcal{P}_i}$ is the identity function for the i-th task. $\mathbb{P}^1$ is the distribution for the generative replay samples drawn from the generator distribution of a GAN model trained on the i-th task learning. The proof is provided in [13].

In order to relieve the forgetting, a GAN model is retrained on its generations in a self-supervised manner in which the generator produces past samples which are mixed with new samples for the new task learning. In this case, the previously generated images and training images from the current task are treated as real samples while the generated images from the generator during the training are treated as fake samples. However, Eq. (90) can only measure the risk of a GAN model for a single target distribution. Inspired by Theorem 1 of the paper, in the following we generalize the expression from (90) to the context of lifelong learning.

Proposition 1. A GAN model is trained on a series of t tasks with the generative replay mechanism. We derive a GB for a GAN model at the t-th task learning :

$$\begin{aligned}\mathcal{E}_{\mathcal{P}_{(1:t)}}\left(h, f_{\mathcal{P}_{(1:t)}}\right) \leq \mathcal{E}_{\mathbb{P}^t}\left(h, f_{\mathcal{P}_{(1:t)}}\right) \\ + \mathcal{L}^{\star}_{\text{disc}}\left(\mathcal{P}_{(1:t)}, \mathbb{P}^t\right) ,\end{aligned} \tag{91}$$

where $\mathcal{P}_{(1:t)}$ represents the distribution that is formed by the real training samples drawn from $\{\mathcal{P}_1, \ldots, \mathcal{P}_t\}$. $\mathbb{P}^t$ is the distribution of the generative samples drawn from the generator of a GAN model which was trained on the t-th task learning. In the following, we derive a GB for a GAN model and show how this model is losing the previously learnt knowledge during the LLL/CL training.

Proposition 2. Let $\{\mathcal{T}_1, \cdots, \mathcal{T}_t\}$ be a sequence of t tasks. We derive a GB for a GAN model at the t-th task learning :

$$\begin{aligned}\sum_{k=1}^{t} \mathcal{E}_{\mathcal{P}_k}\left(h, f_{\mathcal{P}_k}\right) \leq \sum_{k=1}^{t} \Big\{ \mathcal{E}_{\mathbb{P}_k^t}\left(h, f_{\mathcal{P}_k}\right) \\ + \sum_{j=k}^{t} \{\mathcal{L}^{\star}_{\text{disc}}\left(\mathbb{P}_k^{j-1}, \mathbb{P}_k^{j}\right)\} \\ + \mathcal{L}^{\star}_{\text{disc}}\left(\mathbb{P}_k^{k-1}, \mathcal{P}_i\right) \Big\} ,\end{aligned} \tag{92}$$

We provide the detailed proof in Appendix-F from SM. From Eq. (92), we can explicitly evaluate the degenerated performance of a GAN model for each task learning. This is caused by the discrepancy distance term $\mathcal{L}^{\star}_{\text{disc}}(\cdot,\cdot)$ in Eq. (92), which gradually increases while learning more tasks. We can also observe that as the number of tasks t increases, the GAN model would suffer from more forgetting and thus lead to a degenerated performance on the target distributions.

J.2 Energy-based GANs

The Energy-based Generative Adversarial Network (EBGAN), aiming to improve the performance in GANs, was proposed in [14]. Unlike GANs, the discriminator in the EBGAN consists of an auto-encoder that calculates the reconstruction loss as the energy value. During training, EBGAN assigns lower and higher energy values to real and fake data, respectively. We implement $h \in \mathcal{H}$ as the discriminator. The EBGAN model in lifelong learning can be treated as a self-supervised VAE without having the KL divergence term, while its generator produces past samples to relieve forgetting. We derive the following GB for EGBAN when learning the t-th task :

$$\begin{aligned}\frac{1}{t}\sum\nolimits_{i=1}^{t} \left\{\mathcal{E}_{\mathcal{P}_i}\left(h, f_{\mathcal{P}_i}\right)\right\} \leq \mathcal{E}_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\left(h, h^{*}_{\mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t}\right) \\ + \mathcal{E}_A\left(\mathcal{P}_{(1:t)}, \mathbb{P}^{t-1}\otimes\tilde{\mathcal{P}}_t\right) .\end{aligned} \tag{93}$$

Unlike in Theorem 2 of the paper, $\mathbb{P}^{t-1}$ is the distribution of replay samples drawn from the EBGAN generator trained on $(t-1)$ tasks. Since EBGAN does not have the KL regularization, the generator tends to produce more realistic samples when compared to the generator of VAEs. In addition to EBGAN, the proposed theory framework can be used to analyse the forgetting behaviour for other generative models including [15], [16], [17], [18], [19], [20] as well as lifelong learning approaches [3], [21], [22], [23], [24], [25], [26], [27].

REFERENCES

[1] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," in *Proc. of 22nd Conf. on Learning Theory (COLT), arXiv preprint arXiv:0902.3430*, 2009.

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

[3] F. Ye and A. G. Bors, "Learning latent representations across multiple data domains using lifelong VAEGAN," in *Proc. of European Conference on Computer Vision (ECCV), vol. LNCS 12365*, 2020, pp. 777–795.

[4] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," in *Proc. Int. Cont. of Learning Representations (ICLR), arXiv preprint arXiv:1509.00519*, 2015.

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

[6] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[7] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Auxiliary deep generative models," in *Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 48*, 2016, pp. 1445–1453.

[8] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 29, 2016, pp. 4743–4751.

[9] A. Sobolev and D. Vetrov, "Importance weighted hierarchical variational inference," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 601–613.

[10] M. Yin and M. Zhou, "Semi-implicit variational inference," in *Proc. of Int. Conf. on Machine Learning (ICML), vol. PMLR 80*, 2018, pp. 5660–5669.

[11] D. Molchanov, V. Kharitonov, A. Sobolev, and D. Vetrov, "Doubly semi-implicit variational inference," in *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS), vol. PMLR 89*, 2019, pp. 2593–2602.

[12] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, 2010, pp. 2366–2369.

[13] B. Adlam, C. Cortes, M. Mohri, and N. Zhang, "Learning GANs and ensembles using discrepancy," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 5796–5807.

[14] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," in *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1609.03126*, 2017.

[15] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 2016, pp. 2172–2180.

[16] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, "Training generative neural networks via maximum mean discrepancy optimization," in *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2015, pp. 258–267.

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

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

[19] ——, "Deep mixture generative autoencoders," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5789–5803, 2022.

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

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

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

[23] ——, "Lifelong infinite mixture model based on knowledge-driven dirichlet process," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 10 695–10 704.

[24] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," in *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1708.01547*, 2018.

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

[26] ——, "Lifelong teacher-student network learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6280–6296, 2022.

[27] ——, "Lifelong learning of interpretable image representations," in *Proc. Int. Conf. on Image Processing Theory, Tools and Applications (IPTA)*, 2020, pp. 1–6.