Proceedings Paper:
Ye, Fei and Bors, Adrian Gheorghe orcid.org/0000-0001-7838-0021 (2024) Task-Free Dynamic Sparse Vision Transformer for Continual Learning. In: AAAI Conference on Artificial Intelligence. AAAI Press , pp. 16442-16450.

https://doi.org/10.1609/aaai.v38i15.29581

Task-Free Dynamic Sparse Vision Transformer for Continual Learning

Fei Ye[1,2] **and Adrian G. Bors**[1,2]

[1]Department of Computer Science, University of York, York YO10 5GH, UK
[2]Machine Learning Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
fy689@york.ac.uk, adrian.bors@york.ac.uk

Abstract

Vision Transformers (ViTs) represent self-attention-based network backbones shown to be efficient in many individual tasks, but which have not been explored in Task-Free Continual Learning (TFCL) so far. Most existing ViT-based approaches for Continual Learning (CL) are relying on task information. In this study, we explore the advantages of the ViT in a more challenging CL scenario where the task boundaries are unavailable during training. To address this learning paradigm, we propose the Task-Free Dynamic Sparse Vision Transformer (TFDSViT), which can dynamically build new sparse experts, where each expert leverages sparsity to allocate the model's capacity for capturing different information categories over time. To avoid forgetting and ensure efficiency in reusing the previously learned knowledge in subsequent learning, we propose a new dynamic dual attention mechanism consisting of the Sparse Attention (SA') and Knowledge Transfer Attention (KTA) modules. The SA' refrains from updating some previously learned attention blocks for preserving prior knowledge. The KTA uses and regulates the information flow of all previously learned experts for learning new patterns. The proposed dual attention mechanism can simultaneously relieve forgetting and promote knowledge transfer for a dynamic expansion model in a task-free manner. We also propose an energy-based dynamic expansion mechanism using the energy as a measure of novelty for the incoming samples which provides appropriate expansion signals leading to a compact network architecture for TFDSViT. Extensive empirical studies demonstrate the effectiveness of TFDSViT. The code and supplementary material (SM) are available at https://github.com/dtuzi123/TFDSViT.

Introduction

Deep learning models have achieved state-of-the-art performance in many popular vision tasks, including image classification (He et al. 2022), image generation (Goodfellow et al. 2014; Liu, Gu, and Samaras 2019), object detection (Ren et al. 2015) and reconstruction (Kingma and Welling 2013). However, when applying these advanced models for continuously learning a series of tasks, their performance on past tasks would sharply degenerate and eventually they would fail. Learning successively multiple tasks paradigm is called Continual Learning (CL). Classical models tend to rewrite their previously learnt parameters to adapt to new tasks, leading to catastrophic forgetting (Parisi et al. 2019).

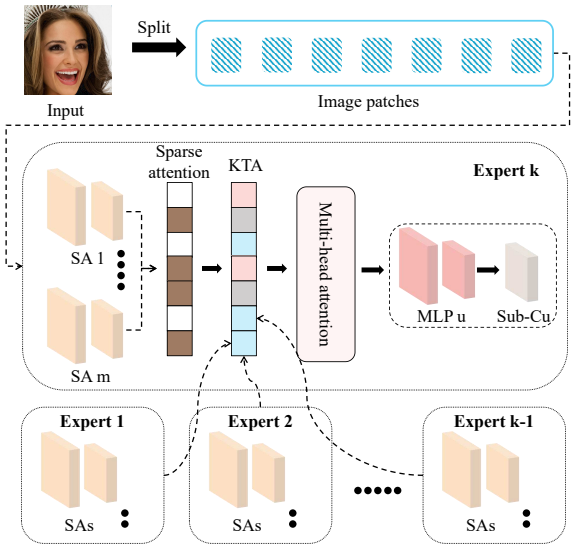


Figure 1: The proposed dynamic dual attention mechanism consists of a Sparse Attention (SA') and a Knowledge Transfer Attention (KTA) module. We assume that TFDSViT has already trained k experts. When learning the u-th submodel consisting of an MLP layer ('MLP u') and a sub-classifier ('Sub-Cu'), the sparse attention generates binary masks for each self-attention block ('SA 1', $\cdots$,'SA m') to avoid forgetting. The KTA generates attention masks to regulate each self-attention block of all previously trained experts incorporating them into the learning process of the u-th submodel. We omit the embedding layer for the sake of simplification.

Existing approaches to continual learning can be broadly divided into three categories : memory-based (Bang et al. 2021, 2022), regularisation-based (Kemker et al. 2018; Martens and Grosse 2015) and dynamic extension models (Hung et al. 2019; Ye and Bors 2020b, 2022b). Memory-based methods store some past examples in a memory buffer, while a classifier is trained on both new and stored examples to relieve forgetting (Bang et al. 2021). Another memory-based approach trains a generator to remember past data and replay generative samples while also learning new

tasks (Zhai et al. 2019; Ye and Bors 2020a). Regularisation-based approaches usually prevent the change of some important parameters to avoid forgetting. These can be combined with the memory-based approaches to further improve their performance (Kemker et al. 2018; Martens and Grosse 2015). However, despite their impressive performance in CL, such methods cannot learn long task sequences due to their fixed model capacity. This drawback prompts several researchers to develop Dynamic Extension Models (DEMs) (Ye and Bors 2021, 2022a), which increase their capacity by adding new models to cope with new tasks. One of the crucial advantages of DEMs over static models is that they can naturally solve the stability-plasticity dilemma by freezing previously learnt parameters and creating new components to learn new data distributions (Ye and Bors 2022a).

The Vision Transformer (ViT) (Dosovitskiy et al. 2021), given its good generalization performance and robust feature learning ability, has recently been used for continual learning (Douillard et al. 2022; Wang et al. 2022a; Xue et al. 2022). However, ViT-based approaches require access to the task information to generate specific attention masks for each learning task (Douillard et al. 2022; Xue et al. 2022), which cannot be considered in a realistic CL scenario where there are no explicit task boundaries (Wang et al. 2022b). Moreover, these approaches are based on either static network architectures (Wang et al. 2022a; Xue et al. 2022) or by employing expansion architectures, where their number of parameters grows ceaselessly while learning an increasing number of tasks (Douillard et al. 2022), which makes them intractable for learning long-term data streams. In this study, we address these challenging problems by proposing a novel ViT-based dynamic expansion model, called the Task-Free Dynamic Sparse Vision Transformer (TFDSViT), which can dynamically maintain a minimum number of sparse experts to deal with emerging patterns without requiring the task information. Each expert consists of multiple self-attention blocks, which are dynamically allocated to capture multiple information categories during different training periods. To appropriately expand the model's capacity, we propose a new dynamic expansion mechanism that estimates the novelty of incoming samples by evaluating their energy similarity. Specifically, we train an auto-encoder as an energy function for each submodel, computing the sample reconstruction error as the energy for a given sample. A high energy score, estimated by all previously learnt autoencoders, indicates that the newly encountered data batch is dissimilar from the already learnt knowledge. Consequently, TFDSViT will allocate or increase its capacities to learn these novel samples. Unlike the task-specific components/tokens used in (Douillard et al. 2022), the proposed dynamic expansion mechanism does not require knowing task boundaries and results in a compact model.

Given that the prior learnt knowledge would be beneficial for future learning, we propose a new dynamic dual attention mechanism that simultaneously relieves forgetting and promotes knowledge transfer for a dynamic expansion model in a task-free manner. Specifically, this mechanism consists of a sparse attention module and a Knowledge Transfer Attention (KTA) module. The former generates binary masks to prevent changes in the parameters of some self-attention blocks, thus preserving previously learnt knowledge while allocating the remaining capacity to learn new samples, as shown in Fig. 1, where we assume that TFDSViT has trained k experts and only the current one ('Expert k') is updated. The KTA module generates attention masks to regulate all previously learnt attention blocks and incorporates them into the learning process of the current expert. These attention masks are continually optimized to minimize the main objective function (classification loss) over time, maximising the positive knowledge transfer benefits.

Our contributions are summarized as follows : (1) We propose a novel ViT-based dynamic expansion model in TFCL, which adaptively expands its capacity without the need for task information. (2) We propose a new dynamic dual attention mechanism which simultaneously prevents catastrophic forgetting and promotes the knowledge transfer for a dynamic expansion model in a task-free manner. (3) We propose an energy-based dynamic expansion mechanism which can dynamically allocate or increase the model's capacities to acquire novel knowledge without any task boundaries. (4) Extensive experiments demonstrate that the proposed TFDSViT far outperforms other methods with less computation or memory costs.

Related Work

Memory-based approaches usually selectively store some past examples using a fixed-length memory buffer, which is used for replaying past information in the subsequent learning to relieve forgetting (Rebuffi et al. 2017; Cha, Lee, and Shin 2021; Tiwari et al. 2022; Wang et al. 2022b,c; Yan et al. 2022). The memory-based approaches can further improve their performance by integrating with the regularization-based models (Li and Hoiem 2017; Dai et al. 2007; Kemker et al. 2018). However, most current memory-based approaches require knowing the task boundaries. The memory-based approaches have been applied to TFCL by developing several sample selection criteria based on the loss value (Aljundi, Kelchtermans, and Tuytelaars 2019; Aljundi et al. 2019a), gradient information (Aljundi et al. 2019b) and a *learner-evaluator* framework (De Lange and Tuytelaars 2021). In addition to the sample selection approach, the Gradient-based Memory EDiting (GMED) (Jin et al. 2021) dynamically modifies the memorized samples, which can then be integrated with memory buffering models to further improve the performance. Although such approaches can achieve promising results, they are not scalable for learning long-term data streams.

The Dynamic Expansion Model (DEM) is a recent popular framework which dynamically expands the network architecture according to the complexity of the given tasks (Hung et al. 2019; Ye and Bors 2020b, 2023b, 2021; Rao et al. 2019; Wen, Tran, and Ba 2020; Ye and Bors 2023a, 2022a, 2023d,c). These approaches preserve the knowledge of past tasks into their frozen parameters while adding new parameters in order to adapt to learning the newly given tasks (Ye and Bors 2021). However, most existing DEMs require knowing the task identity to provide the auxiliary information for the expansion strategy (Ye and Bors 2021). Recently,

DEMs have been shown to achieve promising results in TFCL (Ye and Bors 2022a). The first study of using DEMs for TFCL was proposed in (Rao et al. 2019), which introduces a continual learning framework called the Continual Unsupervised Representation Learning (CURL). CURL can automatically expand its network architecture to adapt to data distribution changes over time. A similar idea called the Continual Neural Dirichlet Process Mixture (CN-DPM) uses the Dirichlet processes for the VAE component expansion (Lee et al. 2020). Moreover DEMs can further improve their performance by using an efficient sample selection approach called the Online Cooperative Memorization (OCM) (Ye and Bors 2022a). which employs a a dual memory system to store both short and long-term information.

The Vision transformer (ViT) : The self-attention mechanism was first used for machine translation in (Vaswani et al. 2017), and was extended to the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019) for language comprehension. Dosovitskiy *et al.* (Dosovitskiy et al. 2021) proposed to split an image into several patches processing them as tokens in a ViT. Several recent works such as the Data-efficient image Transformer (DeiT) (Touvron et al. 2021a), Class-attention in image Transformers (CaiT) (Touvron et al. 2021b), Convolutional vision Transformer (Convit) (d'Ascoli et al. 2021) and the Swin Transformer (Liu et al. 2021) have been proposed to improve the original ViT in terms of computational efficiency and performance. These models can only be applied to a single data set and are not able to handle continuously evolving data distributions. Recently, the DYnamic TOken expansion (DyTox) (Douillard et al. 2022) is the first study to apply ViT to continual learning. DyTox dynamically builds a task-specific trained token when it sees a new task, while sharing most of its parameters across all tasks. However, this approach still requires the task label during the training, which cannot be considered in TFCL. Moreover, DyTox does not control the expansion process, which leads to an ever increasing number of components over time. More discussions are provided in **Appendix-B** from the Supp. Material (SM).

Methodology

Preliminaries

We study a stricter TFCL learning paradigm in which task information and boundaries are unavailable during training and testing. Let us consider a data stream S which consists of incoming samples, assumed to be provided as batches of incremental classes from a training set $D^S = \{\mathbf{x}_i^S, \mathbf{y}_i^S\}_{i=1}^{N^S}$, where $\mathbf{x}_i^S$ and $\mathbf{y}_i^S$ are the observed variable and its class label. N^S represents the total number of training samples. In TFCL, a model can only access a small batch of b samples $\mathbf{X}_e = \{\mathbf{x}_{e,1}^S, \cdots, \mathbf{x}_{e,b}^S\}$ drawn from the data stream S at a certain training time t_e, where b is the batch size. During subsequent learning stages, data batches are drawn from different underlying data distributions, which imposes a severe challenge for the model's learning. The goal is to accumulate knowledge from the learning of the entire data stream and then make accurate predictions on all testing samples.

Overview

The overview of the proposed TFDSViT is presented in Fig. 2. We assume that the proposed TFDSViT has already learnt k experts $\mathbf{E} = \{\mathcal{E}_1, \cdots, \mathcal{E}_k\}$, where each $\mathcal{E}_i$ consists of m self-attention modules. To enable each $\mathcal{E}_i$ to recognize the information from different data categories, we can allow each expert $\mathcal{E}_i$ to dynamically build $u > 1$ Multilayer Perceptrons (MLP) layers and sub-classifiers (Sub-Cs), based on all attention blocks of $\mathcal{E}_i$. Each MLP layer combined with the associated sub-classifier can be seen as a submodel (classifier) for predicting specific concepts/classes.

The proposed dynamic dual attention mechanism consists of a sparse attention (SA') and a Knowledge Transfer Attention (KTA) module. The former aims to allocate the remaining capacity of the current expert in training, during subsequent learning, while preventing the parameter updating of all previously learned submodels. The KTA incorporates and regulates the self-attention blocks of all previously trained experts into the learning procedure of the current expert. KTA automatically generates adaptive attention masks for each previously trained self-attention block, selectively reusing the previously learnt knowledge for learning new patterns, which benefits the positive knowledge transfer, as shown in Fig. 1. Moreover, each expert can dynamically build u auto-encoders, each learning an energy function for the associated submodel. To ensure a compact model structure while promoting the knowledge diversity among submodels for TFDSViT, the proposed dynamic expansion mechanism evaluates the novelty of the incoming samples using all previously trained autoencoders, resulting in appropriate signals for model expansion. Furthermore, these auto-encoders can be employed as model selectors for the submodel, without requiring access to the task information in the testing phase.

The Expert's Architecture

Let $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$ be an input space where $\{H, W, C\}$ denote the image height, weight and channels, respectively. An image $\mathbf{x} \in \mathcal{X}$ is split into N image patches $\mathbf{x}^b = \{\mathbf{x}_1^b, \cdots, \mathbf{x}_N^b\}, \mathbf{x}_i^b \in \mathbb{R}^{G\times G}$, defining local image regions, where G^2 and $N = HW/G^2$ is the patch size and the number of image patches, respectively. Let us define the j-th self-attention block of $\mathcal{E}_k$ as SA_j^k where k represents the expert index. Each self-attention block SA_j^k has a projection matrix $\mathbf{W}_p^{k,j} \in \mathbb{R}^p$ which maps the image patches $\mathbf{x}^b$ into the p-dimensional embedding space :

$$\mathbf{x}_p = \mathbf{W}_p^{k,j}\mathbf{x}^b \,. \tag{1}$$

The embedding vector $\mathbf{x}_p$ is then transformed by each self-attention block SA_j^k which has three trainable weight matrices $\{\mathbf{W}_K^{k,j}, \mathbf{W}_Q^{k,j}, \mathbf{W}_V^{k,j}\}$:

$$\begin{aligned} &\mathcal{S}^{k,j} = \mathrm{Softmax}(\mathbf{Q}^{k,j}(\mathbf{K}^{k,j})^{\mathrm{T}}/\sqrt{d})\mathbf{V}^{k,j}\,, \\ &\mathbf{Q}^{k,j} = \mathbf{W}_Q^{k,j}\mathbf{x}_p, \mathbf{K}^{k,j} = \mathbf{W}_K^{k,j}\mathbf{x}_p, \mathbf{V}^{k,j} = \mathbf{W}_V^{k,j}\mathbf{x}_p\,, \end{aligned} \tag{2}$$

where $\sqrt{d}$ is a scaling factor and $(\cdot)^{\mathrm{T}}$ is the matrix transpose operator. Each expert $\mathcal{E}_k$ has m self-attention blocks

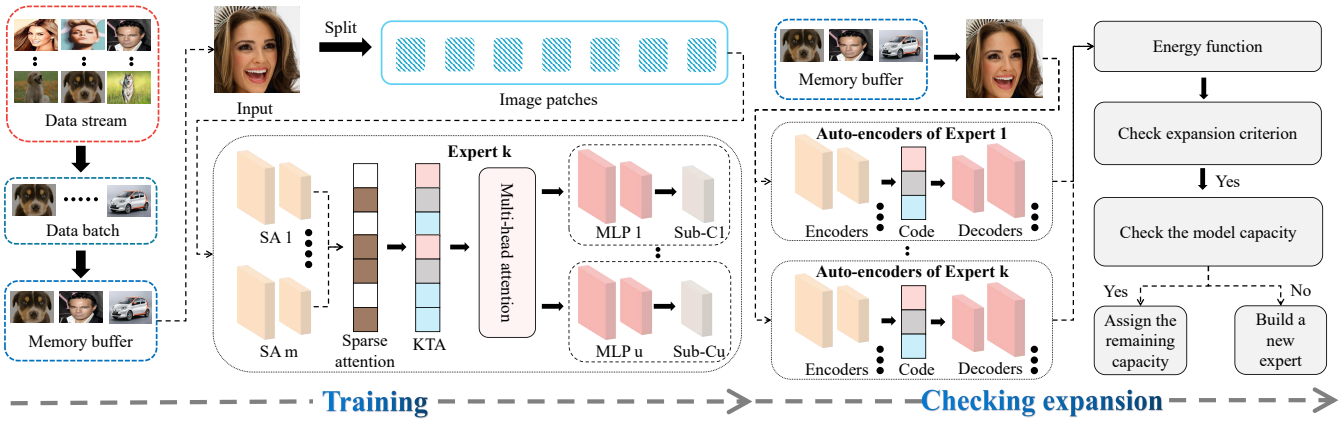


Figure 2: Overview of our proposed model (TFDSViT). We assume that our model has already trained k experts, where each expert is composed of m self-attention (SA) modules and has u MLPs and sub-classifiers. We omit the embedding layer for simplicity. The sparse attention is used to assign the remaining capacity for learning new samples, which allows each expert to capture different category information. In addition, each expert can reuse the information flow from all previously trained experts for a positive knowledge transfer through the knowledge transfer attention. Each expert has u auto-encoders, where each one is assigned to the corresponding submodel. During the training, we only train the current submodel and the associated auto-encoder on the memory buffer. Once the current training step is finished, we check the expansion criterion using Eq. (12).

$\{\mathrm{SA}_1^k, \cdots, \mathrm{SA}_m^k\}$, which can form a multi-head attention mechanism :

$$\mathcal{S}^k = \mathrm{F}_{\mathrm{Concat}}(\mathcal{S}^{k,1}, \cdots, \mathcal{S}^{k,m}), \tag{3}$$

where each $\mathcal{S}^{k,j}, j = 1, \cdots, m$ is obtained using Eq. (2) and $\mathrm{F}_{\mathrm{Concat}}(\cdot)$ is used to concentrate all $\{\mathcal{S}^{k,1}, \cdots, \mathcal{S}^{k,m}\}$. In order to enable each expert $\mathcal{E}_k$ to classify new category information over time, we dynamically create a new submodel $\mathcal{B}^{k,t} = \{\mathrm{F}_{\mathrm{C}}^{k,t}, \mathrm{F}_{\mathrm{MLP}}^{k,t}\}$ consisting of an MLP ($\mathrm{F}_{\mathrm{MLP}}^{k,t}$) and a sub-classifier $\mathrm{F}_{\mathrm{C}}^{k,t}$, on the top of all self-attention blocks $\{\mathrm{SA}_1^k, \cdots, \mathrm{SA}_m^k\}$ from $\mathcal{E}_k$, where the superscript $t = 1, \cdots, u$ represents the submodel index and u is the maximum number of submodels for each expert. For a given input $\mathbf{x}$, the submodel $\mathcal{B}^{k,t}$ makes the prediction $\widehat{p}$:

$$\widehat{p} = \mathrm{F}_{\mathrm{C}}^{k,t}(\mathrm{F}_{\mathrm{MLP}}^{k,t}(\mathcal{S}^k)). \tag{4}$$

We implement the MLP $\mathrm{F}_{\mathrm{MLP}}^{k,t}$ and sub-classifier $\mathrm{F}_{\mathrm{C}}^{k,t}$ using simple fully connected layers to avoid excess parameters.

The Dynamic Dual Attention Mechanism

Current dynamic expansion models (Ye and Bors 2022a) dynamically build a new component when detecting novel information in the data under TFCL. However, they do not consider integrating the already learnt parameters for learning new patterns. In contrast, the proposed TFDSViT enables each expert to dynamically allocate for the model's capacity in order to capture multiple category information at different training periods. This is implemented by the proposed dynamic dual attention mechanism, which consists of SA' and KTA modules. The former generates a binary attention mask vector $\mathbf{W}_M^{k,t} \in \mathbb{R}^m$ for the submodel $\mathcal{B}^{k,t}$ of $\mathcal{E}_k$, which regulates the information flow by :

$$\mathcal{S}^k(t) = \mathrm{F}_{\mathrm{Concat}}(\mathcal{S}^{k,1}\mathbf{W}_M^{k,t}[1], \cdots, \mathcal{S}^{k,m}\mathbf{W}_M^{k,t}[m]), \tag{5}$$

where $\mathbf{W}_M^{k,t}[j]$ denotes the j-th element of $\mathbf{W}_M^{k,t}$ and $\mathbf{W}_M^{k,t}[j] = 0$ indicates that SA_j^k is ignored when learning $\mathcal{B}^{k,t}$, which is ready for future learning use. In practice, we allocate the same capacity for each submodel in an expert.

The KTA module aims to reuse the information flow from all previously trained experts for learning new patterns. Let $\mathbf{W}_A^{k,t} \in \mathbb{R}^{(m(k-1))}$ be a trainable attention vector for $\mathcal{B}^{k,t}$ where $m(t-1)$ represents the number of self-attention blocks of all previously trained experts. To decide which self-attention block contributes more to learning new samples, we propose to normalize $\mathbf{W}_A^{k,t}$ by using the Gumble-Softmax trick (Maddison, Tarlow, and Minka 2014), which also reduces the variation of gradients (Wang et al. 2018) :

$$\widehat{\mathbf{W}}_A^{k,t}[j] = \frac{\exp((\log \mathbf{W}_A^{k,t}[j] + g_j)/T)}{\sum_i^K \exp((\log \mathbf{W}_A^{k,t}[i] + g_i)/T)}, \tag{6}$$

where g_j is a sample drawn from $\mathrm{Gumble}(0, 1)$. $\mathbf{W}_A^{k,t}[j]$ is the j-th element of $\mathbf{W}_A^{k,t}$ and $T = 0.5$ is the temperature parameter. K is the number of self-attention blocks of all previously trained experts. We consider $\widehat{\mathbf{W}}_A^{k,t}$ to regulate each previously learnt self-attention block by :

$$\begin{aligned} \widehat{\mathcal{S}}^{1,1} &= \mathcal{S}^{1,1}\widehat{\mathbf{W}}_A^{k,t}[1], \cdots, \\ \widehat{\mathcal{S}}^{k-1,m} &= \mathcal{S}^{k-1,m}\widehat{\mathbf{W}}_A^{k,t}[m(k-1)]. \end{aligned} \tag{7}$$

Then, we concatenate all weighted self-attention blocks :

$$\begin{aligned} \widehat{\mathcal{S}}^1 &= \mathrm{F}_{\mathrm{Concat}}(\widehat{\mathcal{S}}^{1,1}, \cdots, \widehat{\mathcal{S}}^{1,m}), \cdots \\ \widehat{\mathcal{S}}^{k-1} &= \mathrm{F}_{\mathrm{Concat}}(\widehat{\mathcal{S}}^{k-1,1}, \cdots, \widehat{\mathcal{S}}^{k-1,m}). \end{aligned} \tag{8}$$

An augmented attention map $\widehat{S} = \sum_{i=1}^{k-1}\{\widehat{S}^i\}$ is obtained by considering all previously learnt self-attention blocks using Eq. (8), which concentrates the prior learnt knowledge.

Algorithm 1: Training algorithm for TFDSViT

for $e < n$ **do**
  Memory updating :
  if $|\mathcal{M}_e| > |\mathcal{M}_e|^{max}$ **then**
    Remove the earliest memorized samples
  end if
  $\mathcal{M}_e = \mathcal{M}_{e-1} \bigcup \mathbf{X}_e, \mathbf{X}_e \sim S$ Add a new data batch.
  Training the TFDSViT :
  if $|\mathbf{E}| = 0$ **then**
    $\mathbf{E} = \mathbf{E} \bigcup \mathcal{E}_i$
    $\mathcal{E}_1 = \mathcal{E}_1 \bigcup \mathcal{B}_1$
  else if $|\mathcal{E}_1| = 1$ and $e > |\mathcal{M}_e|^{max}$ **then**
    $\mathcal{E}_1 = \mathcal{E}_1 \bigcup \mathcal{B}_2$ Add a new submodel.
  end if
  Train the current expert $\mathcal{E}_k$ using Eq. (10)
  Train $\mathcal{A}^{k,|\mathcal{E}_k|}$ of $\mathcal{E}_k$ using Eq. (11)
  Check the model expansion :
  if $|\mathcal{M}_e| > |\mathcal{M}_e|^{max}$ **then**
    if $\min\{\mathcal{L}_{\mathrm{AE}}(\mathbf{X}_e, \mathcal{A}^{1,1}), \cdots, \mathcal{L}_{\mathrm{AE}}(\mathbf{X}_e, \mathcal{A}^{k,t-1})\} \geq \gamma$ **then**
      Check the model's capacity :
      if $|\mathcal{E}_k| \geq u$ **then**
        $\mathbf{E} = \mathbf{E} \bigcup \mathcal{E}_{k+1}$
        $\mathcal{E}_{k+1} = \mathcal{E}_{k+1} \bigcup \mathcal{B}^{k+1,1}$
      else
        $\mathcal{E}_k = \mathcal{E}_k \bigcup \mathcal{B}^{k,|\mathcal{E}_k|+1}$
      end if
    end if
  end if
end for

The prediction process of the submodel $\mathcal{B}^{k,t}$ incorporates the augmented attention maps $\widehat{S}$ and $\mathcal{S}^k(t)$:

$$\widehat{p}_c = \mathrm{F}_{\mathrm{C}}^{k,t}(\mathrm{F}_{\mathrm{MLP}}^{k,t}(\mathrm{F}_{\mathrm{Concat}}(\mathcal{S}_c^k(t), \widehat{S}_c))) \,, \tag{9}$$

where $\widehat{p}_c$ is the prediction for $\mathbf{x}_c$, where we use the subscript c to denote that $\mathcal{S}_c^k(t)$ and $\widehat{S}_c$ are obtained using Eq. (5) and Eq. (8) considering $\mathbf{x}_c$. We define a cross-entropy loss function for learning $\mathcal{B}^{k,t}$ at the e-th training time t_e to update the attention and the submodel parameters :

$$\mathcal{L}_{\mathrm{CE}}^{i,t} = \frac{1}{|\mathcal{M}_e|} \sum\nolimits_{c=1}^{|-\mathcal{M}_e|} \sum_{j=1}^{C} \{\mathbf{y}_c[j] \log(\widehat{p}_c[j])\}, \tag{10}$$

where $\mathcal{M}_e$ is a memory buffer updated at t_e and $|\mathcal{M}_e|$ is the total number of memorized samples. $\mathbf{y}_c[j]$ and $\widehat{p}_c[j]$ are the j-th dimension of the c-th class label and the prediction, respectively, and C is the total number of categories. We consider a simple memory updating mechanism removing the earliest memorized samples while storing incoming data.

Energy-Based Dynamic Expansion Mechanism

The energy-based model aims to learn an energy surface in which the data samples consistent with the model have low energies while those classed as outliers are given high energies (LeCun et al. 2006; Zhao, Mathieu, and LeCun 2017). Such a model has been used as a discriminator in adversarial learning (Goodfellow et al. 2014), aiming to distinguish real images from generated ones. Inspired by such energy-based models, we propose to train an autoencoder $\mathcal{A}^{k,t}$ involving an encoder $\mathrm{F}_{\theta^{k,t}}$ and a decoder $\mathrm{F}_{\xi^{k,t}}$ as an energy function for the associated submodel $\mathcal{B}^{k,t}$, where $\theta^{k,t}$ and $\xi^{k,t}$ are the trainable network parameters. The reconstruction error function $\mathcal{L}_{\mathrm{R}}(\cdot,\cdot)$ is used as the energy evaluation as well as the loss function for learning $\mathcal{A}^{k,t}$ at training time t_e :

$$\mathcal{L}_{\mathrm{AE}}(\mathcal{M}_e, \mathcal{A}^{k,t}) = \frac{1}{|\mathcal{M}_e|} \sum_{c=1}^{|\mathcal{M}_e|} \big\{\mathcal{L}_{\mathrm{R}}(\mathbf{x}_c, \mathrm{F}_{\xi^{k,t}}(\mathrm{F}_{\theta^{k,t}}(\mathbf{x}_c)))\big\}. \tag{11}$$

A high energy score, estimated by all previously trained autoencoders, indicates that the incoming data batch is novel and could be used for training a new submodel in order to address the stability-plasticity dilemma in continual learning. To implement this learning paradigm, we propose a novel dynamic expansion criterion which estimates the energy for a new data batch using all previously trained autoencoders :

$$\min\{\mathcal{L}_{\mathrm{AE}}(\mathbf{X}_e, \mathcal{A}^{1,1}), \cdots, \mathcal{L}_{\mathrm{AE}}(\mathbf{X}_e, \mathcal{A}^{k,t-1})\} \geq \gamma, \tag{12}$$

where $\mathbf{X}_e$ is an incoming data batch drawn from the data stream S at the time t_e and γ is an expansion threshold which balances the model size and generalization performance. In addition to fulfilling Eq. (12), we also check whether the current expert $\mathcal{E}_k$ has enough capacity to build a new submodel by :

$$|\mathcal{E}_k| < u \,, \tag{13}$$

where $|\mathcal{E}_k|$ represents the number of current trained submodels for $\mathcal{E}_k$. If Eq. (13) is satisfied, we will build a new submodel $\mathcal{B}^{k,|\mathcal{E}_k|+1}$ by reusing the attention blocks $\{\mathrm{SA}_1^k, \cdots, \mathrm{SA}_m^k\}$, otherwise, we will build a new expert $\mathcal{E}_{k+1}$ into $\mathbf{E} = \{\mathcal{E}_1, \cdots, \mathcal{E}_k, \mathcal{E}_{k+1}\}$.

Algorithm Implementation

This section provides the detailed algorithm implementation while its pseudocode is provided in Algorithm 1. There are three main processing steps at each training time t_e :

- **Step 1 (Memory updating).** At the e-th training time t_e, if the memory buffer is full $|\mathcal{M}_e| = |\mathcal{M}_e|^{max}$, then we remove the earliest memorized samples from $\mathcal{M}_e$, where $|\mathcal{M}_e|^{max}$ is the maximum capacity for the memory buffer. We add a new data batch into the memory buffer, expressed as $\mathcal{M}_e = \mathcal{M}_e \bigcup (\mathbf{X}_e \sim S)$.
- **Step 2 (Training process).** At the training time t_e, if TFDSViT has only one submodel, we preserve the first submodel into $\mathcal{E}_1$ while automatically building the second submodel $\mathcal{B}^{1,2}$ if the memory buffer $\mathcal{M}_e$ is full. Such a mechanism enables the evaluation of the dynamic expansion criterion (Eq. (12)) that requires $\mathbf{E}$ having already trained submodels. We assume that $\mathbf{E}$ has already trained k experts while the current expert $\mathcal{E}_k$ has trained t submodels. During the training, we only optimize the current submodel $\mathcal{B}^{k,t}$ and the associated auto-encoder $\mathcal{A}^{k,t}$ on $\mathcal{M}_e$ using Eq. (10) and Eq. (11), respectively.
- **Step 3 (Checking expansion).** We check the model's expansion (Eq. (12)) when the memory buffer is full $|\mathcal{M}_e| = |\mathcal{M}_e|^{max}$ and $\mathbf{E}$ has trained more than one submodel. We also check the current expert's capacity during the expansion process using Eq. (13). If Eq. (13) is

Methods	**Split MNIST**	**Split CIFAR10**	**Split CIFAR100**
finetune*	19.75 ± 0.05	18.55 ± 0.34	3.53 ± 0.04
GEM*	93.25 ± 0.36	24.13 ± 2.46	11.12 ± 2.48
iCARL*	83.95 ± 0.21	37.32 ± 2.66	10.80 ± 0.37
reservoir*	92.16 ± 0.75	42.48 ± 3.04	19.57 ± 1.79
MIR*	93.20 ± 0.36	42.80 ± 2.22	20.00 ± 0.57
GSS*	92.47 ± 0.92	38.45 ± 1.41	13.10 ± 0.94
CoPE-CE*	91.77 ± 0.87	39.73 ± 2.26	18.33 ± 1.52
CoPE*	93.94 ± 0.20	48.92 ± 1.32	21.62 ± 0.69
ER + GMED†	82.67 ± 1.90	34.84 ± 2.20	20.93 ± 1.60
ER_a + GMED†	82.21 ± 2.90	47.47 ± 3.20	19.60 ± 1.50
WGF-SVGD	-	47.90 ± 2.50	19.90 ± 2.30
CURL*	92.59 ± 0.66	-	-
CNDPM	95.23 ± 0.27	50.26 ± 1.36	28.76 ± 0.57
Dynamic-OCM	95.67 ± 0.22	51.27 ± 1.47	29.87 ± 0.69
TFDSViT	**98.12** ± 0.18	**55.46** ± 1.02	**32.86** ± 0.56

Table 1: Classification accuracy for five independent runs for various models on three datasets. * and † denote the results cited from (De Lange and Tuytelaars 2021) and (Jin et al. 2021), respectively.

Methods	**Split MiniImageNet**
ER_a	25.92 ± 1.2
ER + GMED	27.27 ± 1.8
MIR+GMED	26.50 ± 1.3
MIR	25.21 ± 2.2
CNDPM	27.97 ± 2.3
Dynamic-OCM	28.03 ± 2.1
TFDSViT	**35.62** ± 1.9

Table 2: Classification accuracy for 20 runs when testing various models on Split MiniImageNet.

satisfied, we add a new submodel $\mathcal{B}^{k,|\mathcal{E}_k|+1}$ into $\mathcal{E}_k$, otherwise, we add a new expert $\mathcal{E}_{k+1}$ into $\mathbf{E}$.

Experiments

Experimental Setting

Datasets : We split MNIST (LeCun et al. 1998) containing 60k training samples into five sets and each set has images of two incremental classes (De Lange and Tuytelaars 2021), and call this setting Split MNIST. Similarly, we divide CIFAR10 (Krizhevsky and Hinton 2009) into five sets where each set consists of images from two consecutively ordered classes, named Split CIFAR10. We also split CIFAR100 (Krizhevsky and Hinton 2009) into 20 sets with each set containing images from five incremental classes.

Performance criterion. Since the model in the TFCL scenario does not access the task information, we consider the average classification accuracy on all testing samples as the performance criterion, and this criterion has been used in several other TFCL studies (Aljundi et al. 2019b; De Lange and Tuytelaars 2021; Ye and Bors 2022a).

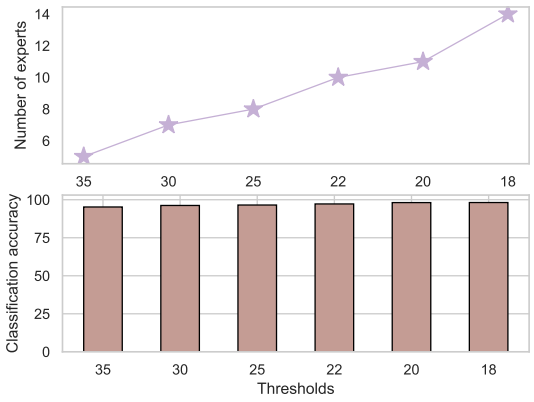
Methods	Split MNIST	Split CIFAR10	Split MImageNet
finetune	21.53 ± 0.1	20.69 ± 2.4	3.05 ± 0.6
ER	79.74 ± 4.0	37.15 ± 1.6	26.47 ± 2.3
MIR	84.80 ± 1.9	38.70 ± 1.7	25.83 ± 1.5
ER + GMED	82.73 ± 2.6	40.57 ± 1.7	28.20 ± 0.6
MIR+GMED	86.17 ± 1.7	41.22 ± 1.1	26.86 ± 0.7
TFDSViT	**93.62** ± 1.3	**49.23** ± 1.2	**32.62** ± 0.5

Table 3: The classification accuracy of five independent runs for various models on fuzzy task boundaries.

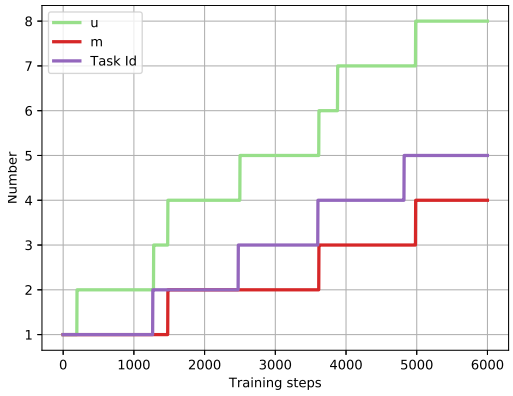
Hyperparameters and implementation. We set the image patch size of 7×7 for Split MNIST. The embedding dimension for Split MNIST is 100. A simple fully connected layer with 100 hidden units implements the MLP module for each submodel. We also implement the encoder and decoder of each autoencoder by using two fully connected layers, with 200 hidden units on each layer. For Split CIFAR10 and Split CIFAR100, we set the image patch size of 8×8 and the embedding dimension as 100. The MLP for each submodel is implemented by two fully connected layers with 500 and 200 hidden units. For all datasets, we consider that each expert has six self-attention blocks $m = 6$ and that we can build two submodels $u = 2$. Additional information is provided in **Appendix-C** from SM.

TFCL Benchmarks

In this section, we evaluate the performance of the proposed TFDSViT on standard TFCL benchmarks. According to the setting from (De Lange and Tuytelaars 2021), a model only sees ten samples once at each training time. The maximum memory buffer size for Split MNIST, Split CIFAR10 and Split CIFAR100 is $|\mathcal{M}_e|^{max} = \{2000, 1000, 5000\}$. The classification accuracy of TFCL benchmarks is reported in Tab. 1, where we compare the proposed TFDSViT with several state-of-the-art methods, including Incremental Classifier and Representation Learning (iCARL) (Rebuffi et al. 2017), GSS (Aljundi et al. 2019b), MIR (Aljundi et al. 2019a), Reservoir (Vitter 1985), Dynamic-Online Cooperative Memorization (OCM) (Ye and Bors 2022a), CURL (Rao et al. 2019), CNDPM (Lee et al. 2020), Gradient Episodic Memory (GEM) (Lopez-Paz and Ranzato 2017), CoPE (De Lange and Tuytelaars 2021), ER + GMED, ER_a + GMED (Jin et al. 2021) where ER is the Experience Replay (ER) (Rolnick et al. 2019) and ER_a denotes that the ER uses data augmentation, WGF-SVGD (Wang et al. 2022b). From Tab. 1, we observe that most dynamic expansion approaches achieve better performance than the static/single model. In addition, the proposed TFDSViT outperforms other methods by a large margin on these three datasets, demonstrating its superior performance. The model complexity (number of



(a) Changing γ in Eq. (12).



(b) Dynamic expansion process.

Figure 3: Ablation study results on Split CIFAR10. (a) The performance and number of experts for TFDSViT when changing γ in Eq. (12). (b) Varying m, u and task ID at each training time.

parameters) of various dynamic expansion models is provided in **Appendix-D5** from SM. These results show that the proposed TFDSViT requires fewer parameters and performs much better than other dynamic expansion models.

In the following we consider the continual learning of Split MiniImageNet which splits MiniImageNet (Vinyals et al. 2016) into twenty sets, where each set contains the images of five consecutive classes (Aljundi et al. 2019a). We employ a CNN network with two convolutional layers as a feature extractor for each expert to reduce the number of parameters. Then the self-attention blocks are built based on the feature extractor. The classification accuracy of various models on Split MiniImageNet is reported in Tab. 2. The proposed TFDSViT achieves better performance than other methods on this more complex dataset. The number of experts trained in the proposed TFDSViT for Split MNIST, Split CIFAR10, Split CIFAR100 and Split MImageNet datasets is of 3, 4, 3, and 4, respectively.

Fuzzy Task Boundaries

In this section, we evaluate the performance of the proposed TFDSViT on a more challenging and realistic CL scenario called fuzzy task learning (Lee et al. 2020). In this setting, samples from the following task are mixed with the samples from the current task after learning half of the current task' data. The classification accuracy of various models on the fuzzy task learning scenario is reported in Tab. 3. From these results, we observe that the proposed TFDSViT achieves the best performance on this challenging CL scenario when compared with other TFCL methods.

Ablation Study

This section investigates the effectiveness of each module in TFDSViT by performing a wide range of ablation studies. More ablation studies are provided in **Appendix-D** from SM.

Effect of the threshold γ : We investigate the model's complexity and generalization performance of the proposed TFDSViT when varying the threshold γ from Eq. (12), and the results on Split MNIST are shown in Fig. 3a. When decreasing γ, the proposed model would dynamically build more experts. In contrast, a large γ prevents the model expansion and induces a compact structure. The result also indicates that by using more experts can not bring a significant performance improvement and TFDSViT with only three experts can capture ten different categories while achieving a good performance as well. These results show that each expert from TFDSViT is able to dynamically allocate its capacity to capture more than one category.

Analysis for the expansion process : We train the proposed TFDSViT on Split MNIST in which we estimate the number of submodels by varying u, of self-attention blocks m for each expert, and the task ID, in each training step in order to investigate the dynamic expansion process. In order to estimate the task information, we assign a task ID for each training sample. Actually, the task ID is not used during the learning process of TFDSViT. The results plotted in Fig. 3b, show that the second expert learns two tasks, demonstrating that each expert in TFDSViT can capture different category information at different training times. The proposed TFDSViT almost always builds a new expert when a new task is given, indicating that it provides good signals for model expansion.

Conclusion

This paper develops the Task-Free Dynamic Sparse Vision Transformer (TFDSViT), which can automatically expand its capacity to adapt to the data distribution shift, without accessing any task information during continual learning. An energy-based dynamic expansion mechanism is proposed to ensure a compact model structure. We then propose a dynamic dual attention mechanism which can simultaneously relieve forgetting and promote knowledge transfer for a dynamic expansion model in a task-free manner, further improving the performance. We evaluate the effectiveness of the proposed TFDSViT against the standard Task-Free Continual Learning (TFCL) baselines and the empirical results show that it outperforms existing methods.

References

Aljundi, R.; Belilovsky, E.; Tuytelaars, T.; Charlin, L.; Caccia, M.; Lin, M.; and Page-Caccia, L. 2019a. Online Continual Learning with Maximal Interfered Retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, 11872–11883.

Aljundi, R.; Kelchtermans, K.; and Tuytelaars, T. 2019. Task-free continual learning. In *Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 11254–11263.

Aljundi, R.; Lin, M.; Goujaud, B.; and Bengio, Y. 2019b. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 11817–11826.

Bang, J.; Kim, H.; Yoo, Y.; Ha, J.-W.; and Choi, J. 2021. Rainbow Memory: Continual Learning with a Memory of Diverse Samples. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 8218–8227.

Bang, J.; Koh, H.; Park, S.; Song, H.; Ha, J.-W.; and Choi, J. 2022. Online Continual Learning on a Contaminated Data Stream With Blurry Task Boundaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9275–9284.

Cha, H.; Lee, J.; and Shin, J. 2021. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9516–9525.

Dai, W.; Yang, Q.; Xue, G. R.; and Yu, Y. 2007. Boosting for transfer learning. In *Proc. Int Conf. on Machine Learning (ICML)*, 193–200.

De Lange, M.; and Tuytelaars, T. 2021. Continual prototype evolution: Learning online from non-stationary data streams. In *Proc. of the IEEE/CVF International Conference on Computer Vision*, 8250–8259.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conf. on Learning Representations (ICLR), arXiv preprint arXiv:2010.11929*.

Douillard, A.; Ramé, A.; Couairon, G.; and Cord, M. 2022. DyTox: Transformers for Continual Learning with DYnamic TOken eXpansion. In *Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 9285–9295.

d'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; and Sagun, L. 2021. ConViT: Improving vision transformers with soft convolutional inductive biases. In *Proc. International Conference on Machine Learning (ICML), vol PMLR 139*, 2286–2296.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2672–2680.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

Hung, C.-Y.; Tu, C.-H.; Wu, C.-E.; Chen, C.-H.; Chan, Y.-M.; and Chen, C.-S. 2019. Compacting, Picking and Growing for Unforgetting Continual Learning. In *Advances in Neural Information Processing Systems*, 13647–13657.

Jin, X.; Sadhu, A.; Du, J.; and Ren, X. 2021. Gradient-based Editing of Memory Examples for Online Task-free Continual Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 29193–29205.

Kemker, R.; McClure, M.; Abitino, A.; Hayes, T.; and Kanan, C. 2018. Measuring catastrophic forgetting in neural networks. In *Proc. of the AAAI Conference on Artificial Intelligence*, 3390–3398.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Univ. of Toronto.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11): 2278–2324.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; and Huang, F. 2006. *Predicting structured data, eds. G. Baklr, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar. S. Vishwanathan*, chapter Energy-Based Models. 10. MIT Press.

Lee, S.; Ha, J.; Zhang, D.; and Kim, G. 2020. A Neural Dirichlet Process Mixture Model for Task-Free Continual Learning. In *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:2001.00689*.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(12): 2935–2947.

Liu, H.; Gu, X.; and Samaras, D. 2019. Wasserstein gan with quadratic transport cost. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4832–4841.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of IEEE/CVF International Conf. on Computer Vision*, 10012–10022.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, 6467–6476.

Maddison, C. J.; Tarlow, D.; and Minka, T. 2014. A* Sampling. *Advances in Neural Inf. Proc. Systems (NIPS)*, 1–9.

Martens, J.; and Grosse, R. B. 2015. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In *Proc. of International Conference on Machine Learning (ICML)*, volume PMLR 37, 2408–2417.

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

Rao, D.; Visin, F.; Rusu, A. A.; Teh, Y. W.; Pascanu, R.; and Hadsell, R. 2019. Continual Unsupervised Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 7645–7655.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. iCaRL: Incremental classifier and representation learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001–2010.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. Advances in Neural Inf. Proc. Systems (NIPS)*, 91–99.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T. P.; and Wayne, G. 2019. Experience Replay for Continual Learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 348–358.

Tiwari, R.; Killamsetty, K.; Iyer, R.; and Shenoy, P. 2022. GCR: Gradient Coreset Based Replay Buffer Selection for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 99–108.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021a. Training data-efficient image transformers & distillation through attention. In *Proc. International Conference on Machine Learning (ICML), vol. PMLR 139*, 10347–10357.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021b. Going deeper with image transformers. In *Proc. of the IEEE/CVF International Conference on Computer Visio (ICCV)*, 32–42.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, 5998–6008.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching networks for one shot learning. *Advances in Neural Information Processing Systems (NIPS)*, 29: 3637–3645.

Vitter, J. S. 1985. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1): 37–57.

Wang, X.; Zhang, R.; Sun, Y.; and Qi, J. 2018. KDGAN: Knowledge Distillation with Generative Adversarial Networks. In *Advances in Neural Infrmation Processing Systems (NeurIPS)*, 775–786.

Wang, Z.; Liu, L.; Duan, Y.; Kong, Y.; and Tao, D. 2022a. Continual Learning With Lifelong Vision Transformer. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 171–181.

Wang, Z.; Shen, L.; Fang, L.; Suo, Q.; Duan, T.; and Gao, M. 2022b. Improving Task-free Continual Learning by Distributionally Robust Memory Evolution. In *International Conference on Machine Learning*, 22985–22998. PMLR.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022c. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Wen, Y.; Tran, D.; and Ba, J. 2020. BatchEnsemble: an Alternative Approach to Efficient Ensemble and Lifelong Learning. In *International Conference on Learning Representations (ICLR), arXiv preprint arXiv:2002.06715*.

Xue, M.; Zhang, H.; Song, J.; and Song, M. 2022. Meta-attention for ViT-backed Continual Learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 150–159.

Yan, Q.; Gong, D.; Liu, Y.; van den Hengel, A.; and Shi, J. Q. 2022. Learning Bayesian Sparse Networks with Full Experience Replay for Continual Learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 109–118.

Ye, F.; and Bors, A. G. 2020a. Learning Latent Representations Across Multiple Data Domains Using Lifelong VAEGAN. In *Proc. European Conf. on Computer Vision (ECCV), vol. LNCS 12365*, 777–795.

Ye, F.; and Bors, A. G. 2020b. Mixtures of variational autoencoders. In *Proc. Int. Conf. on Image Processing Theory, Tools and Applications (IPTA)*, 1–6.

Ye, F.; and Bors, A. G. 2021. Lifelong Infinite Mixture Model Based on Knowledge-Driven Dirichlet Process. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 10695–10704.

Ye, F.; and Bors, A. G. 2022a. Continual Variational Autoencoder Learning via Online Cooperative Memorization. In *Proc. of the European Conference on Computer Vision (ECCV), vol. LNCS 13683*, 531–549.

Ye, F.; and Bors, A. G. 2022b. Dynamic Self-Supervised Teacher-Student Network Learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 45(5): 5731–5748.

Ye, F.; and Bors, A. G. 2023a. Learning dynamic latent spaces for lifelong generative modelling. In *Proc. of AAAI Conference on Artificial Intelligence*, 10891–10899.

Ye, F.; and Bors, A. G. 2023b. Lifelong Mixture of Variational Autoencoders. *IEEE Trans. on Neural Networks and Learning Systems*, 34(1): 461–474.

Ye, F.; and Bors, A. G. 2023c. Self-Evolved Dynamic Expansion Model for Task-Free Continual Learning. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22102–22112.

Ye, F.; and Bors, A. G. 2023d. Wasserstein Expansible Variational Autoencoder for Discriminative and Generative Continual Learning. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 18665–18675.

Zhai, M.; Chen, L.; Tung, F.; He, J.; Nawhal, M.; and Mori, G. 2019. Lifelong GAN: Continual Learning for Conditional Image Generation. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2759–2768.

Zhao, J.; Mathieu, M.; and LeCun, Y. 2017. Energy-based generative adversarial network. In *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1609.03126*.