A Comprehensive Survey of Continual Learning: Theory, Method and Application

Liyuan Wang, Xingxing Zhang, Hang Su, Jun Zhu, *Fellow*, *IEEE*

Abstract—To cope with real-world dynamics, an intelligent system needs to incrementally acquire, update, accumulate, and exploit knowledge throughout its lifetime. This ability, known as continual learning, provides a foundation for AI systems to develop themselves adaptively. In a general sense, continual learning is explicitly limited by catastrophic forgetting, where learning a new task usually results in a dramatic performance degradation of the old tasks. Beyond this, increasingly numerous advances have emerged in recent years that largely extend the understanding and application of continual learning. The growing and widespread interest in this direction demonstrates its realistic significance as well as complexity. In this work, we present a comprehensive survey of continual learning, seeking to bridge the basic settings, theoretical foundations, representative methods, and practical applications. Based on existing theoretical and empirical results, we summarize the general objectives of continual learning as ensuring a proper stability-plasticity trade-off and an adequate intra/inter-task generalizability in the context of resource efficiency. Then we provide a state-of-the-art and elaborated taxonomy, extensively analyzing how representative methods address continual learning, and how they are adapted to particular challenges in realistic applications. Through an in-depth discussion of promising directions, we believe that such a holistic perspective can greatly facilitate subsequent exploration in this field and beyond.

Index Terms—Continual Learning, Incremental Learning, Lifelong Learning, Catastrophic Forgetting.

1 Introduction

Learning is the basis for intelligent systems to accommodate dynamic environments. In response to external changes, evolution has empowered human and other organisms with strong adaptability to continually acquire, update, accumulate and exploit knowledge [150], [229], [328]. Naturally, we expect artificial intelligence (AI) systems to adapt in a similar way. This motivates the study of **continual learning**, where a typical setting is to learn a sequence of contents one by one and behave as if they were observed simultaneously (see Fig. 1, a). Such contents could be new skills, new examples of old skills, different environments, different contexts, etc., with particular realistic challenges incorporated [328], [423]. As the contents are provided incrementally over a lifetime, continual learning is also referred to as **incremental learning** or **lifelong learning** in much of the literature, without a strict distinction [71], [229].

Unlike conventional machine learning models built on the premise of capturing a static data distribution, continual learning is characterized by learning from dynamic data distributions. A major challenge is known as **catastrophic forgetting** [296], [297], where adaptation to a new distribution generally results in a largely reduced ability to capture the old ones. This dilemma is a facet of the trade-off between **learning plasticity** and **memory stability**: an excess of the former interferes with the latter, and vice versa. Beyond simply balancing the "proportions" of these two aspects, a desirable solution for continual learning should obtain strong **generalizability** to accommodate distribution differences within and between tasks (see Fig. 1, b). As a naive baseline, reusing all old training samples (if allowed) makes it easy to address the above challenges, but creates huge computational and storage overheads, as well as potential privacy issues. In fact, continual learning is primarily intended to ensure the **resource efficiency** of model updates, preferably close to learning only new training samples.

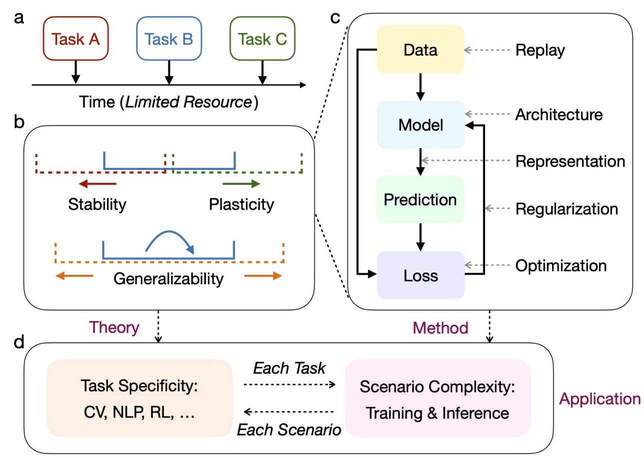


Fig. 1. A conceptual framework of continual learning. **a**, Continual learning requires adapting to incremental tasks with dynamic data distributions (Sec. 2). **b**, A desirable solution should ensure an appropriate trade-off between stability (red arrow) and plasticity (green arrow), as well as an adequate generalizability to intra-task (blue arrow) and inter-task (orange arrow) distribution differences (Sec. 3). **c**, To achieve the objective of continual learning, representative methods have targeted various aspects of machine learning (Sec. 4). **d**, Continual learning is adapted to practical applications to address particular challenges such as scenario complexity and task specificity (Sec. 5).

A number of continual learning methods have been proposed in recent years for various aspects of machine learning, which can be conceptually separated into five groups (see Fig. 1, c): adding regularization terms with

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu are with Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Center, THBI Lab, Tsinghua-Bosch Joint Center for ML, Tsinghua University, Beijing, China (email: wly19@tsinghua.org.cn; xxzhang1993@gmail.com; {suhangss, dcszj}@tsinghua.edu.cn). Corresponding author: Jun Zhu.

reference to the old model (regularization-based approach); approximating and recovering the old data distributions (replay-based approach); explicitly manipulating the optimization programs (optimization-based approach); learning robust and well-distributed representations (representation-based approach); and constructing task-adaptive parameters with a properly-designed architecture (architecture-based approach). This taxonomy extends the commonly-used ones with current advances, and provides refined sub-directions for each category. We summarize how these methods achieve the objective of continual learning, with an extensive analysis of their theoretical foundations and specific implementations. In particular, these methods are *closely connected*, e.g., regularization and replay ultimately act to rectify the gradient directions in optimization, and *highly synergistic*, e.g., the efficacy of replay can be facilitated by distilling knowledge from the old model.

Realistic applications present particular challenges for continual learning, categorized into *scenario complexity* and *task specificity* (see Fig. 1, d). As for the former, for example, the task identity is probably missing in training and testing, and the training samples might be introduced in tiny batches or even one pass. Due to the expense and scarcity of data labeling, continual learning needs to be effective for few-shot, semi-supervised and even unsupervised scenarios. As for the latter, although current advances mainly focus on visual classification, other vision domains such as object detection and semantic segmentation, as well as other related fields, such as conditional generation, reinforcement learning (RL), natural language processing (NLP) and ethic considerations, are receiving increasing attention with their own characteristics. We summarize their particular challenges and analyze how continual learning methods are adapted to them.

Considering the significant growth of interest in continual learning, we believe that such an *up-to-date* and *comprehensive* survey can provide a holistic perspective for subsequent work. Although there are some early surveys of continual learning with relatively broad coverage [71], [328], important advances in recent years have not been incorporated. In contrast, the latest surveys typically capture only a partial aspect of continual learning, with respect to its biological underpinnings [150], [157], [187], [229], specialized settings for visual classification [86], [215], [288], [294], [354], as well as specific extensions for NLP [38], [209] or RL [214]. To the best of our knowledge, this is the first survey that systematically summarizes the latest advances in continual learning. Building on such strengths, we provide an in-depth discussion of continual learning, in terms of current trends, cross-directional prospects and interdisciplinary connections with neuroscience.

The paper is organized as follows: In Sec. 2, we introduce the setups of continual learning, including its basic formulation, typical scenarios and evaluation metrics. In Sec. 3, we summarize the theoretical efforts on continual learning in response to its general objectives, which motivate the development of various continual learning methods. In Sec. 4, we present a state-of-the-art and elaborated taxonomy of representative methods, analyzing their motivations and typical implementations. In Sec. 5 and 6, we describe how these methods are adapted to realistic applications in terms of scenario complexity and task specificity. In Sec. 7, we further discuss current trends, cross-directional prospects and interdisciplinary connections with neuroscience.

2 Setup

In this section, we first present a basic formulation of continual learning. Then we introduce typical scenarios and evaluation metrics.

2.1 Basic Formulation

Continual learning is characterized as learning from dynamic data distributions. In practice, training samples of different distributions arrive in sequence. A continual learning model parameterized by θ needs to learn corresponding task(s) with no or limited access to old training samples and perform well on their test sets. Formally, an incoming batch of training samples belonging to a task t can be represented as $\mathcal{D}_{t,b} = \{\mathcal{X}_{t,b}, \mathcal{Y}_{t,b}\}$, where $\mathcal{X}_{t,b}$ is the input data, $\mathcal{Y}_{t,b}$ is the data label, $t \in \mathcal{T} = \{1, \cdots, k\}$ is the task identity and $b \in \mathcal{B}_t$ is the batch index ($\mathcal{T}$ and $\mathcal{B}_t$ denote their space, respectively). Here we define a "task" by its training samples $\mathcal{D}_t$ following the distribution $\mathbb{D}_t := p(\mathcal{X}_t, \mathcal{Y}_t)$ ($\mathcal{D}_t$ denotes the entire training set by omitting the batch index, likewise for $\mathcal{X}_t$ and $\mathcal{Y}_t$), and assume that there is no difference in distribution between training and testing. Under realistic constraints, the data label $\mathcal{Y}_t$ and the task identity t might not be always available. In continual learning, the training samples of each task can arrive incrementally in batches (i.e., $\{\{\mathcal{D}_{t,b}\}_{b\in\mathcal{B}_t}\}_{t\in\mathcal{T}}$) or simultaneously (i.e., $\{\mathcal{D}_t\}_{t\in\mathcal{T}}$).

2.2 Typical Scenario

According to the division of incremental batches and the availability of task identities, we describe typical continual learning scenarios as follows (please refer to Table 1 for a formal comparison):

- *Instance-Incremental Learning* (IIL): All training samples belong to the same task and arrive in batches.
- *Domain-Incremental Learning* (DIL): Tasks have the same data label space but different input distributions. Task identities are not required.
- *Task-Incremental Learning* (TIL): Tasks have disjoint data label spaces. Task identities are provided in both training and testing.
- *Class-Incremental Learning* (CIL): Tasks have disjoint data label spaces. Task identities are only provided in training.
- *Task-Free Continual Learning* (TFCL): Tasks have disjoint data label spaces. Task identities are not provided in either training or testing.
- *Online Continual Learning* (OCL): Tasks have disjoint data label spaces. Training samples for each task arrive as a one-pass data stream.
- *Blurred Boundary Continual Learning* (BBCL): Task boundaries are blurred, characterized by distinct but overlapping data label spaces.
- *Continual Pre-training* (CPT): Pre-training data arrives in sequence. The goal is to improve knowledge transfer to downstream tasks.

TABLE 1
A formal comparison of typical continual learning scenarios. $\mathcal{D}_{t,b}$: the training samples of task t and batch b. $|b|$: the size of batch b. $\mathcal{B}_t$: the space of incremental batches belonging to task t. $\mathcal{D}_t$: the training set of task t (further specified as $\mathcal{D}_t^{pt}$ for pre-training). $\mathcal{T}$: the space of all incremental tasks (further specified as $\mathcal{T}^{pt}$ for pre-training). $\mathcal{X}_t$: the input data in $\mathcal{D}_t$. $p(\mathcal{X}_t)$: the distribution of $\mathcal{X}_t$. $\mathcal{Y}_t$: the data label of $\mathcal{X}_t$.

Scenario	Training	Testing
IIL [279]	$\{\{\mathcal{D}_{t,b}, t\}_{b\in\mathcal{B}_t}\}_{t=j}$	$\{p(\mathcal{X}_t)\}_{t=j}$; t is not required
DIL [167], [423]	$\{\mathcal{D}_t, t\}_{t\in\mathcal{T}}$; $p(\mathcal{X}_i) \neq p(\mathcal{X}_j)$ and $\mathcal{Y}_i = \mathcal{Y}_j$ for $i \neq j$	$\{p(\mathcal{X}_t)\}_{t\in\mathcal{T}}$, t is not required
TIL [167], [423]	$\{\mathcal{D}_t, t\}_{t\in\mathcal{T}}$; $p(\mathcal{X}_i) \neq p(\mathcal{X}_j)$ and $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$ for $i \neq j$	$\{p(\mathcal{X}_t)\}_{t\in\mathcal{T}}$; t is available
CIL [167], [423]	$\{\mathcal{D}_t, t\}_{t\in\mathcal{T}}$; $p(\mathcal{X}_i) \neq p(\mathcal{X}_j)$ and $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$ for $i \neq j$	$\{p(\mathcal{X}_t)\}_{t\in\mathcal{T}}$; t is unavailable
TFCL [15]	$\{\{\mathcal{D}_{t,b}\}_{b\in\mathcal{B}_t}\}_{t\in\mathcal{T}}$; $p(\mathcal{X}_i) \neq p(\mathcal{X}_j)$ and $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$ for $i \neq j$	$\{p(\mathcal{X}_t)\}_{t\in\mathcal{T}}$; t is optionally available
OCL [16]	$\{\{\mathcal{D}_{t,b}\}_{b\in\mathcal{B}_t}\}_{t\in\mathcal{T}}$, $|b| = 1$; $p(\mathcal{X}_i) \neq p(\mathcal{X}_j)$ and $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$ for $i \neq j$	$\{p(\mathcal{X}_t)\}_{t\in\mathcal{T}}$; t is optionally available
BBCL [27], [46]	$\{\mathcal{D}_t, t\}_{t\in\mathcal{T}}$; $p(\mathcal{X}_i) \neq p(\mathcal{X}_j)$, $\mathcal{Y}_i \neq \mathcal{Y}_j$ and $\mathcal{Y}_i \cap \mathcal{Y}_j \neq \emptyset$ for $i \neq j$	$\{p(\mathcal{X}_t)\}_{t\in\mathcal{T}}$; t is unavailable
CPT [409]	$\{\mathcal{D}_t^{pt}, t\}_{t\in\mathcal{T}^{pt}}$, followed by a downstream task j	$\{p(\mathcal{X}_t)\}_{t=j}$; t is not required

If not specified, each task is often assumed to have a sufficient number of labeled training samples, i.e., *Supervised Continual Learning*. According to the provided $\mathcal{X}_t$ and $\mathcal{Y}_t$ in each $\mathcal{D}_t$, continual learning is further extended to zero-shot [134], [397], few-shot [414], semi-supervised [440], open-world (i.e., to identify unknown classes and then incorporate their labels) [194], [445] and un-/self-supervised [168], [363] scenarios. Besides, other practical challenges have been considered and incorporated, such as multiple labels [216], noisy labels [28], [217], hierarchical granularity [2] and sub-populations [257], mixture of task similarity [210], long-tailed distribution [271], domain alignment [259], domain shifting [395], anytime inference [226], novel class discovery [195], [377], multi-modality [402], etc. Some recent work has focused on various combinations of these scenarios [28], [49], [226], [279], [303], [473], as a way to better simulate the complexity of the real world.

2.3 Evaluation Metric

In general, the performance of continual learning can be evaluated from three aspects: overall performance of the tasks learned so far, memory stability of old tasks, and learning plasticity of new tasks. Here we summarize the common evaluation metrics, using classification as an example.

Overall performance is typically evaluated by *average accuracy* (AA) [63], [281] and *average incremental accuracy* (AIA) [102], [165]. Let $a_{k,j} \in [0, 1]$ denote the classification accuracy evaluated on the test set of the j-th task after incremental learning of the k-th task ($j \leq k$). The output space to compute $a_{k,j}$ consists of the classes in either $\mathcal{Y}_j$ or $\cup_{i=1}^k \mathcal{Y}_i$, corresponding to the use of multi-head evaluation (e.g., TIL) or single-head evaluation (e.g., CIL) [63]. The two metrics at the k-th task are then defined as

$$\mathrm{AA}_k = \frac{1}{k} \sum_{j=1}^{k} a_{k,j}, \tag{1}$$

$$\mathrm{AIA}_k = \frac{1}{k} \sum_{i=1}^{k} \mathrm{AA}_i, \tag{2}$$

where AA represents the overall performance at the current moment and AIA further reflects the historical performance.

Memory stability can be evaluated by *forgetting measure* (FM) [63] and *backward transfer* (BWT) [281]. As for the former, the forgetting of a task is calculated by the difference between its maximum performance obtained in the past and its current performance:

$$f_{j,k} = \max_{i\in\{1,...,k-1\}} (a_{i,j} - a_{k,j}), \forall j < k. \tag{3}$$

FM at the k-th task is the average forgetting of all old tasks:

$$\mathrm{FM}_k = \frac{1}{k-1} \sum_{j=1}^{k-1} f_{j,k}. \tag{4}$$

As for the latter, BWT evaluates the average influence of learning the k-th task on all old tasks:

$$\mathrm{BWT}_k = \frac{1}{k-1} \sum_{j=1}^{k-1} (a_{k,j} - a_{j,j}), \tag{5}$$

where the forgetting is usually reflected as a negative BWT.

Learning plasticity can be evaluated by *intransience measure* (IM) [63] and *forward transfer* (FWT) [281]. IM is defined as the inability of a model to learn new tasks, calculated by the difference of a task between its joint training performance and continual learning performance:

$$\mathrm{IM}_k = a_k^* - a_{k,k}, \tag{6}$$

where a_k^* is the classification accuracy of a randomly-initialized reference model jointly trained with $\cup_{j=1}^k \mathcal{D}_j$ for the k-th task. In comparison, FWT evaluates the average influence of all old tasks on the current k-th task:

$$\mathrm{FWT}_k = \frac{1}{k-1} \sum_{j=2}^{k} (a_{j,j} - \tilde{a}_j), \tag{7}$$

where $\tilde{a}_j$ is the classification accuracy of a randomly-initialized reference model trained with $\mathcal{D}_j$ for the j-th task. Note that, $a_{k,j}$ can be adapted to other forms depending on the task type, such as average precision (AP) for object detection [393], Intersection-over-Union (IoU) for semantic segmentation [56], Fréchet Inception Distance (FID) for image generation [464], normalized reward for reinforcement learning [9], etc, and the above evaluation metrics should be adjusted accordingly.

Besides, there are many other useful metrics, such as linear probes for representation forgetting [84], maximum eigenvalue of the Hessian matrix for flatness of loss landscape [313], area under the curve of accuracy for anytime inference [226], as well as the overheads of storage and computation for resource efficiency [66], etc. We refer readers to their original papers.

3 Theoretical Foundation

In this section, we summarize the theoretical efforts on continual learning with respect to both stability-plasticity trade-off and generalizability analysis, and relate them to the motivations of various continual learning methods.

3.1 Stability-Plasticity Trade-off

With the basic formulation in Sec. 2.1, let's consider a general setup for continual learning, where a neural network with parameters $\theta \in \mathbb{R}^{|\theta|}$ needs to learn k incremental tasks. The training set and test set of each task are assumed to follow the same distribution $\mathbb{D}_t$, $t = 1, ..., k$, where the training set $\mathcal{D}_t = \{\mathcal{X}_t, \mathcal{Y}_t\} = \{(x_{t,n}, y_{t,n})\}_{n=1}^{N_t}$ includes N_t data-label pairs. The objective is to learn a probabilistic model $p(\mathcal{D}_{1:k}|\theta) = \prod_{t=1}^{k} p(\mathcal{D}_t|\theta)$ (with assumption of conditional independence) that can perform well on all tasks denoted as $\mathcal{D}_{1:k} := \{\mathcal{D}_1, ..., \mathcal{D}_k\}$. The task-related performance for discriminative models can be expressed as $\log p(\mathcal{D}_t|\theta) = \sum_{n=1}^{N_t} \log p_\theta(y_{t,n}|x_{t,n})$. The central challenge of continual learning generally arises from the sequential nature of learning: when learning the k-th task from $\mathcal{D}_k$, the old training sets $\{\mathcal{D}_1, ..., \mathcal{D}_{k-1}\}$ are inaccessible. Therefore, it is critical but difficult to capture the distributions of both old and new tasks in a balanced manner, i.e., ensuring an appropriate **stability-plasticity trade-off**, where excessive learning plasticity or memory stability can largely compromise each other (see Fig. 2, a, b).

A straightforward idea is to approximate and recover the old data distributions by storing a few old training samples or training a generative model, known as the *replay-based approach* in Sec. 4.2. According to the learning theory for supervised learning [154], the performance of an old task is improved with replaying more old training samples that approximate its distribution, but resulting in potential privacy issues and a linear increase in **resource overhead**. The use of generative models is also limited by a huge additional resource overhead, as well as their own catastrophic forgetting and expressiveness.

An alternative choice is to propagate the old data distributions in updating parameters through formulating continual learning in a Bayesian framework. Based on a prior $p(\theta)$ of the network parameters, the posterior after observing the k-th task can be computed with Bayes' rule:

$$p(\theta|\mathcal{D}_{1:k}) \propto p(\theta) \prod_{t=1}^{k} p(\mathcal{D}_t|\theta) \propto p(\theta|\mathcal{D}_{1:k-1}) p(\mathcal{D}_k|\theta), \quad (8)$$

where the posterior $p(\theta|\mathcal{D}_{1:k-1})$ of the (k-1)-th task becomes the prior of the k-th task, and thus enables the new posterior $p(\theta|\mathcal{D}_{1:k})$ to be computed with only the current training set $\mathcal{D}_k$. However, as the posterior is generally intractable (except very special cases), a common option is to approximate it with $q_{k-1}(\theta) \approx p(\theta|\mathcal{D}_{1:k-1})$, likewise for $q_k(\theta) \approx p(\theta|\mathcal{D}_{1:k})$. In the following, we will introduce two widely-used approximation strategies:

The first is online *Laplace approximation*, which approximates $p(\theta|\mathcal{D}_{1:k-1})$ as a multivariate Gaussian with local gradient information [177], [202], [222], [371], [441]. Specifically, we can parameterize $q_{k-1}(\theta)$ with ϕ_{k-1} and construct an approximate Gaussian posterior $q_{k-1}(\theta) := q(\theta; \phi_{k-1}) = \mathcal{N}(\theta; \mu_{k-1}, \Lambda_{k-1}^{-1})$ through performing a second-order Taylor expansion around the mode $\mu_{k-1} \in \mathbb{R}^{|\theta|}$ of $p(\theta|\mathcal{D}_{1:k-1})$, where Λ_{k-1} denotes the precision matrix and $\phi_{k-1} = \{\mu_{k-1}, \Lambda_{k-1}\}$, likewise for $q(\theta; \phi_k)$, μ_k and Λ_k. According to Eq. (8), the posterior mode for learning the current k-th task can be computed as

$$\begin{aligned}\mu_k &= \arg\max_\theta \log p(\theta|\mathcal{D}_{1:k}) \\ &\approx \arg\max_\theta \log p(\mathcal{D}_k|\theta) + \log q(\theta; \phi_{k-1}) \\ &= \arg\max_\theta \log p(\mathcal{D}_k|\theta) - \frac{1}{2}(\theta - \mu_{k-1})^\top \Lambda_{k-1}(\theta - \mu_{k-1}),\end{aligned} \quad (9)$$

which is updated recursively from μ_{k-1} and Λ_{k-1}. Meanwhile, Λ_k is updated recursively from Λ_{k-1}:

$$\begin{aligned}\Lambda_k &= -\nabla_\theta^2 \log p(\theta|\mathcal{D}_{1:k})\big|_{\theta=\mu_k} \\ &\approx -\nabla_\theta^2 \log p(\mathcal{D}_k|\theta)\big|_{\theta=\mu_k} + \Lambda_{k-1},\end{aligned} \quad (10)$$

where the first term on the right side is the Hessian of the negative log likelihood of $\mathcal{D}_k$ at μ_k, denoted as $H(\mathcal{D}_k, \mu_k)$. In practice, $H(\mathcal{D}_k, \mu_k)$ is often computationally inefficient due to the great dimensionality of $\mathbb{R}^{|\theta|}$, and there is no guarantee that the approximated Λ_k is positive semi-definite for the Gaussian assumption. To overcome these issues, the Hessian is usually approximated by the Fisher information matrix (FIM):

$$F_k = \mathbb{E}[\nabla_\theta \log p(\mathcal{D}_k|\theta) \nabla_\theta \log p(\mathcal{D}_k|\theta)^\top]\big|_{\theta=\mu_k} \approx H(\mathcal{D}_k, \mu_k). \quad (11)$$

For ease of computation, the FIM can be further simplified with a diagonal approximation [177], [222] or a Kronecker-factored approximation [293], [371]. Then, Eq. (9) is implemented by saving a frozen copy of the old model μ_{k-1} to regularize parameter changes, known as the *regularization-based approach* in Sec. 4.1. Here, we use EWC [222] as an example and present its loss function:

$$\mathcal{L}_{\text{EWC}}(\theta) = \ell_k(\theta) + \frac{\lambda}{2}(\theta - \mu_{k-1})^\top \hat{F}_{1:k-1}(\theta - \mu_{k-1}), \quad (12)$$

where ℓ_k denotes the task-specific loss, the FIM $\hat{F}_{1:k-1} = \sum_{t=1}^{k-1} \text{diag}(F_t)$ with a diagonal approximation $\text{diag}(\cdot)$ of each F_t, and λ is a hyperparameter to control the strength of regularization.

The second is online *variational inference* (VI) [7], [91], [206], [235], [248], [280], [318], [378], [410]. There are many different ways to do this, and a representative one is to minimize the following KL-divergence over a family $\mathcal{Q}$ that satisfies $p(\theta|\mathcal{D}_{1:k}) \in \mathcal{Q}$ at the current k-th task:

$$q_k(\theta) = \arg\min_{q \in \mathcal{Q}} \text{KL}(q(\theta) \parallel \frac{1}{Z_k} q_{k-1}(\theta) p(\mathcal{D}_k|\theta)), \quad (13)$$

where Z_k is the normalizing constant of $q_{k-1}(\theta)p(\mathcal{D}_k|\theta)$. In practice, the above minimization can be achieved by employing an additional Monte Carlo approximation, with specifying $q_k(\theta) := q(\theta;\phi_k) = \mathcal{N}(\theta;\mu_k,\Lambda_k^{-1})$ as a multivariate Gaussian. Here we use VCL [318] as an example, which minimizes the following objective (i.e., maximize its negative):

$$\mathcal{L}_{\mathrm{VCL}}(q_k(\theta)) = \mathbb{E}_{q_k(\theta)}(\ell_k(\theta)) + \mathrm{KL}(q_k(\theta) \parallel q_{k-1}(\theta)), \quad (14)$$

where the KL-divergence can be computed in a closed-form and serves as an implicit regularization term. In particular, although the loss functions of Eq. (12) and Eq. (14) take similar forms, the former is a local approximation at a set of deterministic parameters θ, while the latter is computed by sampling from the variational distribution $q_k(\theta)$. This is attributed to the fundamental difference between the two approximation strategies [318], [420], with slightly different performance in particular settings.

In addition to the parameter space, the idea of sequential Bayesian inference is also applicable to the function space [326], [378], [417], which tends to enable more flexibility in updating parameters. Also, there are many other extensions of VI, such as improving posterior updates with variational auto-regressive Gaussian processes (VAR-GPs) [206], constructing task-specific parameters [7], [232], [244], [280], and adapting to a non-stationary data stream [235].

In essence, the constraint on continual learning for either replay or regularization is ultimately reflected in gradient directions. As a result, some recent work directly manipulates the gradient-based optimization process, categorized as the *optimization-based approach* in Sec. 4.3. Specifically, when a few old training samples $\mathcal{M}_t$ for task t are maintained in a memory buffer, gradient directions of the new training samples are encouraged to stay close to that of the $\mathcal{M}_t$ [66], [281], [411]. This is formulated as $\langle \nabla_\theta \mathcal{L}_k(\theta;\mathcal{D}_k), \nabla_\theta \mathcal{L}_k(\theta;\mathcal{M}_t)\rangle \geq 0$ for $t \in \{1,...,k-1\}$, so as to essentially enforce non-increase in the loss of old tasks, i.e., $\mathcal{L}_k(\theta;\mathcal{M}_t) \leq \mathcal{L}_k(\theta_{k-1};\mathcal{M}_t)$, where θ_{k-1} is the network parameters at the end of learning the (k-1)-th task.

Alternatively, gradient projection can also be performed without storing old training samples [65], [115], [146], [202], [227], [258], [266], [333], [380], [448], [496]. Here we take NCL [202] as an example, which manipulates gradient directions with μ_{k-1} and Λ_{k-1} in online Laplace approximation. As shown in Eq. (15), NCL performs continual learning by minimizing the task-specific loss $\ell_k(\theta)$ within a region of radius r centered around θ with a distance metric $d(\theta,\theta+\delta) = \sqrt{\delta^\top\Lambda_{k-1}\delta/2}$ that takes into account the curvature of the prior via its precision matrix Λ_{k-1}:

$$\begin{aligned} \delta^* &= \arg\min_\delta \ell_k(\theta+\delta) \\ &\approx \arg\min_\delta \ell_k(\theta) + \nabla_\theta \ell_k(\theta)^\top \delta, \\ &\text{s.t.}, d(\theta,\theta+\delta) = \sqrt{\delta^\top\Lambda_{k-1}\delta/2} \leq r. \end{aligned} \quad (15)$$

The solution to such an optimization problem in Eq. (15) is given by $\delta^* \propto \Lambda_{k-1}^{-1}\nabla_\theta \ell_k(\theta) - (\theta-\mu_{k-1})$, which gives rise to the following update rule for a learning rate λ:

$$\theta \leftarrow \theta + \lambda[\Lambda_{k-1}^{-1}\nabla_\theta \ell_k(\theta) - (\theta-\mu_{k-1})], \quad (16)$$

in which the first term encourages parameter changes predominantly in directions that do not interfere with the old tasks via a preconditioner Λ_{k-1}^{-1}, while the second term enforces θ to stay close to the old task solution μ_{k-1}.

Of note, the above analyses are mainly based on finding a shared solution for all incremental tasks, which is subject to severe inter-task interference [361], [441], [443]. In contrast, incremental tasks can also be learned in a (partially) separated way, which is the dominant idea of the *architecture-based approach* in Sec. 4.5. This can be formulated as constructing a continual learning model with parameters $\theta = \cup_{t=1}^k \theta^{(t)}$, where $\theta^{(t)} = \{e^{(t)},\psi\}$, $e^{(t)}$ is the task-specific/adaptive parameters, and ψ is the task-sharing parameters. The task-sharing parameters ψ are omitted in some cases, where the task-specific parameters $e^{(i)}$ and $e^{(j)}$ ($i<j$) may overlap to enable parameter reuse and knowledge transfer. The overlapping part $e^{(i)} \cap e^{(j)}$ is frozen when learning the j-th task to avoid catastrophic forgetting [379], [384]. Then, each task can be performed as $p(\mathcal{D}_t|\theta^{(t)})$ instead of $p(\mathcal{D}_t|\theta)$ if given the task identity $\mathbb{I}_{\mathcal{D}_t}$, in which the conflicts between tasks are explicitly controlled or even avoided if ψ is omitted:

$$\begin{aligned} p(\mathcal{D}_t|\theta) &= \sum_{i=1}^k p(\mathcal{D}_t|\mathbb{I}_{\mathcal{D}_t}=i,\theta)p(\mathbb{I}_{\mathcal{D}_t}=i|\theta) \\ &= p(\mathcal{D}_t|\mathbb{I}_{\mathcal{D}_t}=t,\theta)p(\mathbb{I}_{\mathcal{D}_t}=t|\mathcal{D}_t,\theta) \\ &= p(\mathcal{D}_t|\theta^{(t)})p(\mathbb{I}_{\mathcal{D}_t}=t|\mathcal{D}_t,\theta) \\ &= p(\mathcal{D}_t|e^{(t)},\psi)p(\mathbb{I}_{\mathcal{D}_t}=t|\mathcal{D}_t,\theta). \end{aligned} \quad (17)$$

However, there are two major challenges. The first is the scalability of model size due to progressive allocation of $\theta^{(t)}$, which depends on the sparsity of $e^{(t)}$, reusability of $e^{(i)} \cap e^{(j)}$ ($i<j$), and transferability of ψ. The second is the accuracy of task-identity prediction, denoted as $p(\mathbb{I}_{\mathcal{D}_t}=t|\mathcal{D}_t,\theta)$. Except for the TIL setting that always provides the task identity $\mathbb{I}_{\mathcal{D}_t}$ [108], [117], [379], [384], other scenarios generally require the model to determine which $\theta^{(t)}$ to use based on the input data, as shown in Eq. (17). This is closely related to the out-of-distribution (OOD) detection, where the predictive uncertainty should be low for in-distribution data and high for OOD data [83], [161], [218]. More importantly, since the function of task-identity prediction as Eq. (18) (equivalent to classifying tasks) needs to be continually updated, it also suffers from catastrophic forgetting. To address this issue, the i-th task's distribution $p(\mathcal{D}_t|i,\theta)$ can be recovered by incorporating replay [139], [161], [223], [244]:

$$p(\mathbb{I}_{\mathcal{D}_t}=i|\mathcal{D}_t,\theta) \propto p(\mathcal{D}_t|i,\theta)p(i), \quad (18)$$

where the marginal task distribution $p(i) \propto N_i$ in general.

3.2 Generalizability Analysis

Current theoretical efforts for continual learning have primarily been performed on training sets of incremental tasks, assuming that their test sets follow similar distributions and the candidate solutions have similar generalizability. However, since the objective for learning multiple tasks is typically highly non-convex, there are many local optimal solutions that perform similarly on each training set but have significantly different generalizability on test sets [313], [443]. For continual learning, a desirable solution requires not only *intra-task generalizability* from training sets to test sets, but also *inter-task generalizability* to accommodate to accommodate incremental changes of their distributions.

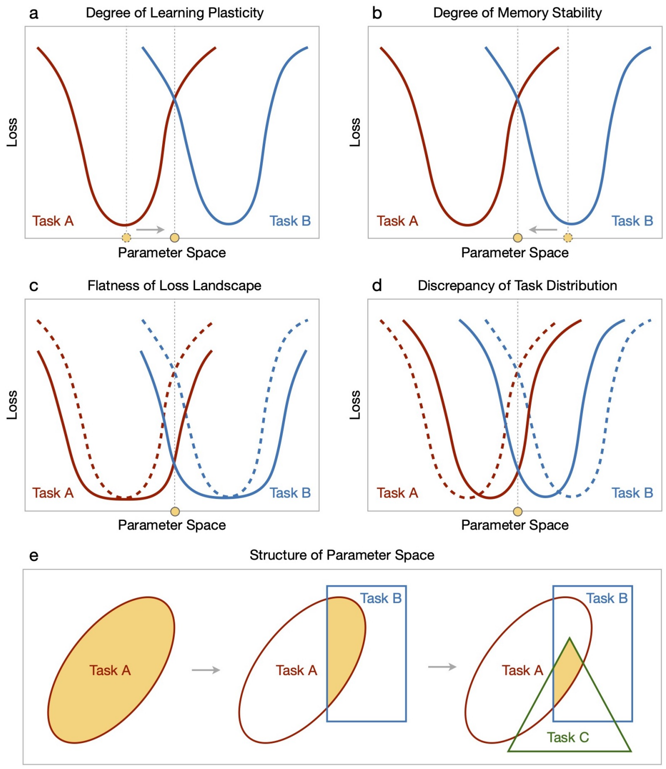


Fig. 2. Analysis of critical factors for continual learning. **a**, **b**, Continual learning requires a proper balance between learning plasticity and memory stability, where excess of either can affect the overall performance (adapted from [443]). **c**, **d**, When the converged loss landscape is flatter and the observed data distributions are more similar, a properly balanced solution can better generalize to the task sequence (adapted from [443]). **e**, The structure of parameter space determines the complexity and possibility of finding a desirable solution (adapted from [225]). The yellow area indicates the feasible parameter space shared by individual tasks, which tends to be narrow and irregular as more incremental tasks are introduced.

Here we provide a conceptual demonstration with a task-specific loss $\ell_t(\theta; \mathcal{D}_t)$ and its empirical optimal solution $\theta_t^* = \arg\min_\theta \ell_t(\theta; \mathcal{D}_t)$. When a task i needs to accommodate another task j, the maximum sacrifice of its performance can be estimated by performing a second-order Taylor expansion of $\ell_i(\theta; \mathcal{D}_i)$ around θ_i^*:

$$\begin{aligned}\ell_i(\theta_j^*; \mathcal{D}_i) &\approx \ell_i(\theta_i^*; \mathcal{D}_i) + (\theta_j^* - \theta_i^*)^\top \nabla_\theta \ell_i(\theta; \mathcal{D}_i)\big|_{\theta=\theta_i^*} \\ &+ \frac{1}{2}(\theta_j^* - \theta_i^*)^\top \nabla_\theta^2 \ell_i(\theta; \mathcal{D}_i)\big|_{\theta=\theta_i^*}(\theta_j^* - \theta_i^*) \\ &\approx \ell_i(\theta_i^*; \mathcal{D}_i) + \frac{1}{2}\Delta\theta^\top \nabla_\theta^2 \ell_i(\theta; \mathcal{D}_i)\big|_{\theta=\theta_i^*}\Delta\theta,\end{aligned} \tag{19}$$

where $\Delta\theta := \theta_j^* - \theta_i^*$ and $\nabla_\theta \ell_i(\theta; \mathcal{D}_i)\big|_{\theta=\theta_i^*} \approx \mathbf{0}$. Then, the performance degradation of task i is upper-bounded by

$$\ell_i(\theta_j^*; \mathcal{D}_i) - \ell_i(\theta_i^*; \mathcal{D}_i) \leq \frac{1}{2}\lambda_i^{max}\|\Delta\theta\|^2, \tag{20}$$

where λ_i^{max} is the maximum eigenvalue of the Hessian matrix $\nabla_\theta^2 \ell_i(\theta; \mathcal{D}_i)\big|_{\theta=\theta_i^*}$. Note that the order of task i and j can be arbitrary, that is, Eq. (20) demonstrates both forward and backward effects. Therefore, the robustness of an empirical optimal solution θ_i^* to parameter changes is closely related to λ_i^{max}, which has been a common metric to describe the **flatness of loss landscape** [163], [212], [313].

Intuitively, convergence to a local minima with a flatter loss landscape will be less sensitive to modest parameter changes and thus benefit both old and new tasks (see Fig. 2, c). To find such a *flat minima*, there are three widely-used strategies in continual learning. The first is derived from its definition, i.e., the flatness metric. Specifically, the minimization of $\ell_t(\theta; \mathcal{D}_t)$ can be replaced by a robust task-specific loss $\ell_t^b(\theta; \mathcal{D}_t) := \max_{\|\delta\| \leq b} \ell_t(\theta + \delta; \mathcal{D}_t)$, thus the obtained solution guarantees low error not only at a specific point but also in its neighborhood with a "radius" of b. However, due to the great dimensionality of θ, the calculation of $\ell_t^b(\theta; \mathcal{D}_t)$ cannot cover all possible δ but only a few directions [387], similar to the complexity issue of computing the Hessian matrix in Eq. (19). The alternatives include using an approximation of the Hessian [90], [313] or calculating δ only along the trajectory of forward and backward parameter changes [58], [174], [183], [300], [312]. The second is to operate the loss landscape by constructing an ensemble model under the restriction of mode connectivity, i.e., integrating multiple minima in parameter or function space along the low-error path, since connecting them ensures flatness on that path [60], [131], [183], [312], [443], [462]. These two strategies are closely related to the *optimization-based approach*. The third comes down to obtaining well-distributed representations, which tend to be more robust to distribution differences in function space, such as by using pre-training [168], [300], [360], wider network architectures [309], [359], [360] and self-supervised learning [57], [168], [286], [343]. Observing the substantial attention to large-scale pre-training and self-supervised learning, we group this direction into the *representation-based approach* in Sec. 4.4.

There are many other factors that are important for continual learning performance. As shown in Eq. (20), the upper bound of performance degradation also depends on the difference of the empirical optimal solution $\theta_t^* = \arg\min_\theta \ell_t(\theta; \mathcal{D}_t)$ for each task, i.e., the **discrepancy of task distribution** (see Fig. 2, d), which is further validated by a theoretical analysis of the forgetting-generalization trade-off [358] and the PAC-Bayes bound of generalization errors [338], [443]. Therefore, how to exploit task similarity is directly related to the performance of continual learning. The generalization error for each task can improve with synergistic tasks but deteriorate with competing tasks [361], where learning all tasks equally in a shared solution tends to compromise each task in performance [361], [443].

On the other hand, when model parameters are *not* shared by all tasks (e.g., using a multi-head output layer), the impact of task similarity on continual learning will be complex. Some theoretical studies with the neural tangent kernel (NTK) [33], [95], [207], [243] suggest that an increase in task similarity may lead to more forgetting. Since the output heads are independent for individual tasks, it becomes much more difficult to distinguish between two similar solutions [207], [243]. Specifically, under the NTK regime from the t-th task up until the k-th task, the forgetting of old tasks is bounded by:

$$\begin{aligned}&\|p(\mathcal{D}_k|\theta_k^*) - p(\mathcal{D}_k|\theta_t^*)\|_F^2 \leq \\ &\sigma_{t,|\mathrm{rep}|+1}^2 \sum_{i=t+1}^{k} \left\|\Theta^{t\to S(i,|\mathrm{rep}|)}\right\|_2^2 \left\|\Theta^{i\to S(i,|\mathrm{rep}|)}\right\|_2^2 \|\mathrm{RES_i}\|_2^2.\end{aligned} \tag{21}$$

$\Theta^{t\to k}$ is a diagonal matrix where each diagonal element

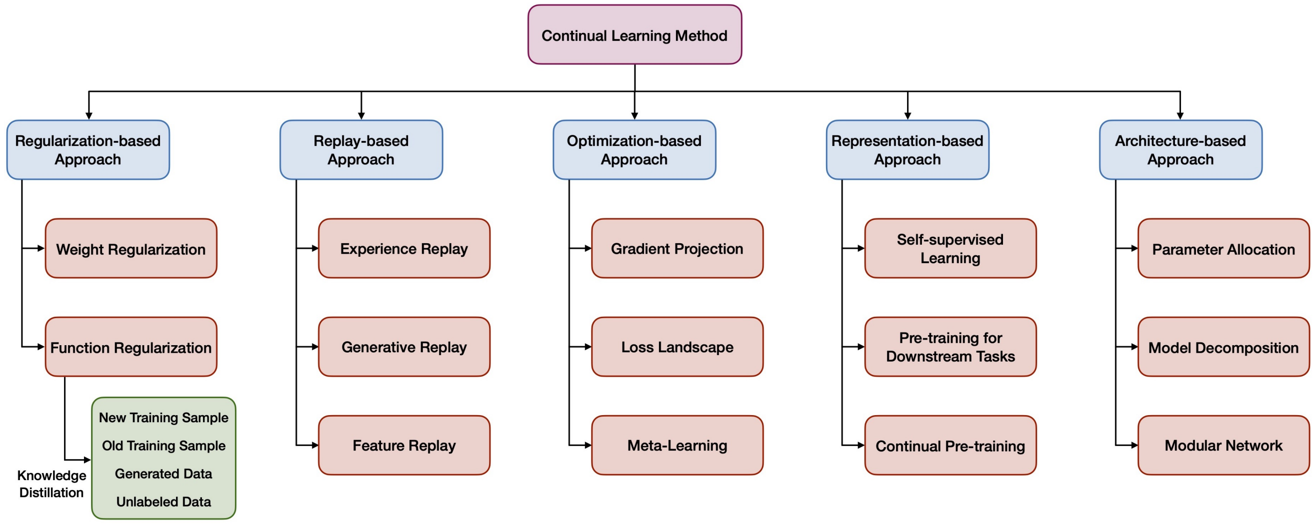


Fig. 3. A state-of-the-art and elaborated taxonomy of representative continual learning methods. We have summarized five main categories (blue blocks), each of which is further divided into several sub-directions (red blocks).

$\cos(\theta_{t,k})_r$ is the cosine of the r-th principal angle between the t-th and k-th tasks in the feature space. $\sigma_{t,\cdot}$ is the $\cdot$-th singular value of the t-th task. RES_i is the rotated residuals that remain to be learned, and $S(i,\cdot)$ represents the residuals subspace of order $\cdot$ until the i-th task. $|\mathrm{rep}|$ is the sample number of replay data. The complex impact of task similarity suggests the importance of model architectures for coordinating task-sharing and task-specific components.

Moreover, the complexity of finding a desirable solution for continual learning is determined to a large extent by the **structure of parameter space**. Learning all incremental tasks with a shared solution is equivalent to learning each new task in a constrained parameter space that prevents performance degradation of all old tasks. Such a classical continual learning problem has proven to be NP-hard in general [225], because the feasible parameter space tends to be narrow and irregular as more tasks are introduced, thus difficult to identify (see Fig. 2, e). This challenging issue can be mitigated by replaying representative old training samples [225], restricting the feasible parameter space to a hyperrectangle [459], or alternating the model architecture of using a single parameter space (e.g., using multiple continual learning models) [96], [361], [443].

To harmonize the important factors in continual learning, recent work presents a similar form of generalized bounds for learning and forgetting. For example, with probability at least $1-\delta$ for any $\delta \in (0,1)$, an ideal continual learner [337] under the assumption that all tasks share a global minimizer with uniform convergence, i.e., $\lambda_i^{max} = \lambda$ for $\forall t = 1, \cdots, k$ in Eq. (20), has the generalization bound

$$c_t^* \leq \mathbb{E}_{\mathcal{D}_t \sim \mathbb{D}_t} \ell_t(\theta; \mathcal{D}_t) \leq c_t^* + \zeta(N_t, \delta), \forall t = 1, \cdots, k, \quad (22)$$

where $c_t^* = \ell_t(\theta_t^*; \mathcal{D}_t)$ is the minimum loss of the t-th task, and θ is a global solution of the continually learned $1:k$ tasks by empirical risk minimization. $\zeta = O(\frac{\lambda B \sqrt{|\theta| \log(N_t) \log(|\theta| k/\delta)}}{2\sqrt{N_t}})$, and $\|\theta\|_2 \leq B$. Considering that the shared parameter space for many different tasks might be an empty set (see Fig. 2, e), i.e., $\cup_{t=1}^k \theta_t = \emptyset$, the generalization bounds are further refined by assuming K parameter spaces ($K \geq 1$ in general) to capture all tasks [442], [443]. For generalization errors of new and old tasks:

$$\mathbb{E}_{\mathcal{D}_t \sim \mathbb{D}_t} \ell_t(\theta; \mathcal{D}_t) \leq c_t^* + R(\sum_{i=1}^{t-1} \ell_i^b) + \sum_{i=1}^{t-1} \mathrm{Div}(\mathbb{D}_i, \mathbb{D}_t) + \zeta(\sum_{\mathrm{i}=1}^{t-1} \mathrm{N_i}, \mathrm{K}/\delta),$$
$$\sum_{i=1}^{t-1} \mathbb{E}_{\mathcal{D}_i \sim \mathbb{D}_i} \ell_i(\theta; \mathcal{D}_i) \leq \sum_{i=1}^{t-1} c_i^* + R(\ell_t^b) + \sum_{i=1}^{t-1} \mathrm{Div}(\mathbb{D}_t, \mathbb{D}_i) + \zeta(\mathrm{N_t}, \mathrm{K}/\delta), \quad (23)$$

where $R(\cdot)$ and Div are the functions of loss flatness and task discrepancy, respectively. The definitions of δ, θ and c_t^* are the same as Eq. (22).

These theoretical efforts suggest that, a desirable solution for continual learning should provide an appropriate stability-plasticity trade-off and an adequate intra/inter-task generalizability, motivating a variety of representative methods as detailed in the next section.

4 METHOD

In this section, we present an elaborated taxonomy of representative continual learning methods (see Fig. 3 and also Fig. 1, c), analyzing extensively their main motivations, typical implementations and empirical properties.

4.1 Regularization-Based Approach

This direction is characterized by adding explicit regularization terms to balance the old and new tasks, which usually requires storing a frozen copy of the old model for reference (see Fig. 4). Depending on the target of regularization, such methods can be divided into two sub-directions.

The first is **weight regularization**, which selectively regularizes the variation of network parameters. A typical implementation is to add a quadratic penalty in loss function that penalizes the variation of each parameter depending on its contribution or "importance" to performing the old tasks, e.g., Eq. (12), in a form originally derived from online Laplace approximation of the posterior under the Bayesian framework. The importance can be calculated by the Fisher information matrix (FIM), such as EWC [222] and some more advanced versions [371], [382]. Meanwhile, numerous efforts have been devoted to designing a better importance measurement. SI [497] online approximates the importance of each parameter by its contribution to the total loss variation and its update length over the entire training trajectory. MAS [12] accumulates the importance measurements based on the sensitivity of predictive results to parameter changes, which is both online and unsupervised. RWalk [63] combines the regularization terms of SI [497] and EWC [222] to integrate their advantages. Interestingly, these importance measurements have been shown to be all tantamount to an approximation of the FIM [34], although stemming from different motivations.

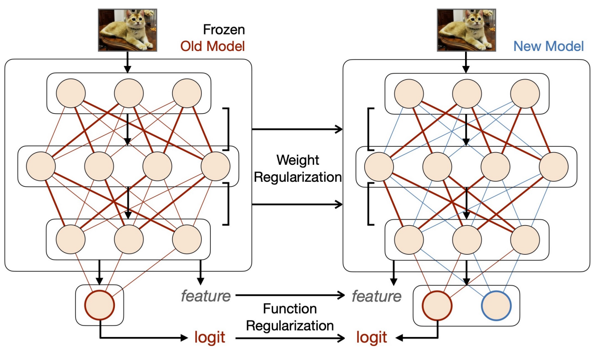


Fig. 4. Regularization-based approach. This direction is characterized by adding explicit regularization terms to mimic the parameters (weight regularization) or behaviors (function regularization) of the old model.

There are also several works refining the implementation of the quadratic penalty. Since the diagonal approximation of the FIM in EWC might lose information about the old tasks, R-EWC [272] performs a factorized rotation of the parameter space to diagonalize the FIM. XK-FAC [239] further considers the inter-example relations in approximating the FIM to better accommodate batch normalization. Observing the asymmetric effect of parameter changes on old tasks, ALASSO [329] designs an asymmetric quadratic penalty with one of its sides overestimated.

Compared to learning each task within the constraints of the old model, which typically exacerbates the intransience, an *expansion-renormalization* process of obtaining separately the new task solution and renormalizing it with the old model has been shown to provide a better stability-plasticity trade-off [240], [245], [258], [382], [441]. IMM [245] is an early attempt that incrementally matches the moment of the posterior distributions of old and new tasks, i.e., a weighted average of their solutions. ResCL [240] extends this idea with a learnable combination coefficient. P&C [382] learns each task individually with an additional network, and then distills it back to the old model with a generalized version of EWC. AFEC [441] introduces a forgetting rate to eliminate the potential negative transfer from the original posterior $p(\theta|\mathcal{D}_{1:k-1})$ in Eq. (8), and derives quadratic terms to penalize differences of the network parameters θ with both the old and new task solutions. To reliably average the old and new task solutions, a linear connector [258] is constructed by constraining them on a linear low-error path.

Other forms of regularization that target the network itself also belong to this sub-direction. As discussed before, online VI of the posterior distribution can serve as an implicit regularization of parameter changes, such as VCL [318], [410], NVCL [420], CLAW [7], GVCL [280], KCL [91] and VAR-GPs [206]. Instead of consolidating parameters, NPC [325] estimates the importance of each neuron and selectively reduces its learning rate. UCL [9] and AGS-CL [196] freeze the parameters connecting the important neurons, equivalent to a hard version of weight regularization.

The second is **function regularization**, which targets the intermediate or final output of the prediction function. This strategy typically employs the previously-learned model as the teacher and the currently-trained model as the student, while implementing knowledge distillation (KD) [141] to mitigate catastrophic forgetting. Ideally, the target of KD should be all old training samples, which are unavailable in continual learning. The alternatives can be new training samples [93], [181], [255], [362], a small fraction of old training samples [53], [102], [165], [365], external unlabeled data [241], generated data [464], [500], etc., suffering from different degrees of distribution shift.

As a pioneer work, LwF [255] and LwF.MC [365] learn *new training samples* while using their predictions from the output head of the old tasks to compute the distillation loss. LwM [93] exploits the attention maps of new training samples for KD. EBLL [362] learns task-specific autoencoders and prevents changes in feature reconstruction. GD [241] further distills knowledge from massive *unlabeled data* available in the wild. When old training samples are faithfully recovered, the potential of function regularization can be largely released. Thus, function regularization often collaborates with replaying a few *old training samples*, such as iCaRL [365], EEIL [53], LUCIR [165], PODNet [102], DER [46], etc., discussed latter in Sec. 4.2. Besides, sequential Bayesian inference over function space can be seen as a form of function regularization, which generally requires storing some old training samples (called "coreset" in literature), such as FRCL [417], FROMP [326] and S-FSVI [378]. For conditional generation, the *generated data* of previously-learned conditions and their output values are regularized to be consistent between the old and new models, such as MeRGANs [464], DRI [452] and LifelongGAN [500].

4.2 Replay-Based Approach

We group the methods for approximating and recovering old data distributions into this category (see Fig. 5). Depending on the content of replay, these methods can be further divided into three sub-directions, each with its own targets and challenges.

The first is **experience replay**, which typically stores a few old training samples within a small memory buffer. Due to the extremely limited storage space, the key challenges consist of *how to construct* and *how to exploit* the memory buffer. As for construction, the preserved old training samples should be carefully selected, compressed, augmented, and updated, in order to recover adaptively the past information. Earlier work adopts fixed principles for *sample selection*. For example, Reservoir Sampling [67], [369], [430] randomly preserves a fixed number of old training samples obtained from each training batch. Ring Buffer [281] further ensures an equal number of old training samples per class. Mean-of-Feature [365] selects an equal number of old training samples that are closest to the feature mean of each class. There are many other fixed principles, such as k-means [67], plane distance [369] and entropy [369], but all perform mediocrely [67], [369]. More advanced strategies are typically gradient-based or optimizable, by maximizing such as the sample diversity in terms of parameter gradients (GSS [16]), performance of corresponding tasks with cardinality constraints (CCBO [40]), mini-batch gradient similarity and cross-batch gradient diversity (OCS [491]), ability of optimizing latent decision boundaries (ASER [391]), diversity of robustness against perturbations (RM [27]), similarity to the gradients of old training samples with respect to the current parameters (GCR [418]), etc.

To improve *storage efficiency*, AQM [48] performs online continual compression based on a VQ-VAE framework [424] and saves compressed data for replay. MRDC [444] formulates experience replay with data compression as determinantal point processes (DPPs) [231], and derives a computationally efficient way for online determination of an appropriate compression rate. RM [27] adopts conventional and label mixing-based strategies of data augmentation to enhance the diversity of old training samples. RAR [234] synthesizes adversarial samples near the forgetting boundary and performs MixUp [504] for data augmentation. The auxiliary information with low storage cost, such as class statistics (IL2M [31], SNCL [139]) and attention maps (RRR [110], EPR [380]), can be further incorporated to maintain old knowledge. Besides, the old training samples could be continually modified to accommodate incremental changes, such as by making them more representative (Mnemonics [276]) or more challenging (GMED [192]) for separation.

As for exploitation, experience replay requires an adequate use of the memory buffer to recover the past information. There are many different implementations, closely related to other continual learning strategies. First, the effect of old training samples in *optimization* can be constrained to avoid catastrophic forgetting and facilitate knowledge transfer. For example, GEM [281] constructs individual constraints based on the old training samples for each task to ensure non-increase in their losses. A-GEM [66] replaces the individual constraints with a global loss of all tasks to improve training efficiency. LOGD [411] decomposes the gradient of each task into task-sharing and task-specific components to leverage inter-task information. To achieve a good trade-off in interference-transfer [369] (i.e., forgetting-generalization [358]), MER [369] employs meta-learning for gradient alignment in experience replay. BCL [358] explicitly pursues a saddle point of the cost of old and new training samples. MetaSP [407] leverages the Pareto optimum of example influence on stability-plasticity to control the model and storage updates. To selectively utilize the memory buffer, MIR [13] prioritizes the old training samples that subject to more forgetting, while HAL [64] uses them as "anchors" and stabilizes their predictions.

On the other hand, experience replay can be naturally combined with *knowledge distillation* (KD), which additionally incorporates the past information from the old model. iCaRL [365] and EEIL [53] are two early works that perform

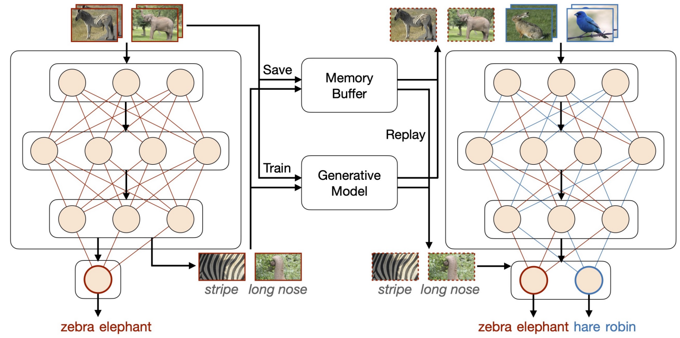


Fig. 5. Replay-based approach. This direction is characterized by approximating and recovering the old data distributions. Typical sub-directions include experience replay, which saves a few old training samples in a memory buffer; generative replay, which trains a generative model to provide generated samples; and feature replay, which recovers the distribution of old features through saving prototypes, saving statistical information or training a generative model.

KD on both old and new training samples. Some subsequent improvements focus on different issues in experience replay. To mitigate data imbalance of the limited old training samples, LUCIR [165] encourages similar feature orientation of the old and new models, while performing cosine normalization of the last layer and mining hard negatives of the current task. BiC [468] and WA [512] attribute this issue to the bias of the last fully connected layer, and resolve it by either learning a bias correction layer with a balanced validation set [468] or normalizing the output weights [512]. SS-IL [10] adopts separated softmax in the last layer and task-wise KD to mitigate the bias. DRI [452] trains a generative model to supplement the old training samples with additional generated data. To alleviate dramatic representation shifts, PODNet [102] employs a spatial distillation loss to preserve representations throughout the model. Co2L [57] introduces a self-supervised distillation loss to obtain robust representations against catastrophic forgetting. GeoDL [396] performs KD along a path that connects the low-dimensional projections of the old and new feature spaces for a smooth transition between them. ELI [193] learns an energy manifold with the old and new models to realign the representation shifts for optimizing incremental tasks. To adequately exploit the past information, AU [236] incorporates uncertainty and self-attention into the distillation loss, while CSC [21] additionally leverages the structure of the feature space. DDE [172] distills colliding effects from the features of the new training samples, which is causally equivalent to replaying more old training samples. TAMiL [36] adds task-specific attention in the feature space and performs consistency regularization to better preserve task-relevant information. To further enhance learning plasticity, D+R [164] performs KD from an additional model dedicated to each new task. FOSTER [435] dynamically expands new modules to fit the residuals of the old model and then distills them into a single model. Besides, weight regularization approaches can be combined with experience replay to achieve better performance and generality [63], [441].

It is worth noting that the merits and potential limitations of experience replay remain largely open. In addition to the intuitive benefits of staying in the low-loss region of the old tasks [428], theoretical analysis has demonstrated its contribution to resolving the NP-hard problem of optimal continual learning [225]. However, it risks overfitting to only a few old training samples retained in the memory buffer, which potentially affects generalizability [428]. To alleviate overfitting, LiDER [39] takes inspirations from adversarial robustness and enforces the Lipschitz continuity of the model to its inputs. MOCA [493] enlarges the variation of representations to prevent the old ones from shrinking in their space. Interestingly, several simple baselines of experience replay can achieve considerable performance. DER/DER++ [46] and X-DER [41] preserve old training samples together with their logits, and perform logit-matching throughout the optimization trajectory. GDumb [348] greedily collects incoming training samples in a memory buffer and then uses them to train a model from scratch for testing. These efforts can serve as an evaluation criterion for more advanced strategies in this sub-direction.

The second is **generative replay** or pseudo-rehearsal, which usually requires training an additional generative model to replay generated data. This is closely related to continual learning of generative models themselves, as they also require incremental updates. DGR [392] provides an initial framework in which learning each generation task is accompanied with replaying generated data sampled from the old generative model, so as to inherit the previously-learned knowledge. MeRGAN [464] further employs replay alignment to enforce consistency of the generative data sampled with the same random noise between the old and new generative models, similar to the role of function regularization. Besides, other continual learning strategies can be incorporated into generative replay. Weight regularization [318], [383], [438], [440] and experience replay [158], [440] have been shown to be effective in mitigating catastrophic forgetting of generative models. DGMa/DGMw [322] adopts binary masks to allocate task-specific parameters for overcoming inter-task interference, and an extendable network to ensure scalability. If pre-training is available, it can provide a relatively stable and strong reference model for continual learning. For example, FearNet [211] and ILCAN [472] additionally preserves statistical information of the old features acquired from a pre-trained feature extractor, while GAN-Memory [81] continually adjusts a pre-trained generative model with task-specific parameters.

The generative models for pseudo-rehearsal can be of various types, such as generative adversarial networks (GANs) and (variational) autoencoder (VAE). A majority of state-of-the-art approaches have focused on GANs to enjoy its advantages in fine-grained generation, but usually suffer from label inconsistency in continual learning [25], [322]. In contrast, autoencoder-based strategies, such as FearNet [211], SRM [370], CLEER [375], EEC [25], GMR [342] and Flashcards [140], can explicitly control the labels of the generated data, albeit with relatively blurred granularity. L-VAEGAN [483] instead employs a hybrid model for both high-quality generation and accurate inference. However, since continual learning of generative models is extremely difficult and requires significant resource overhead, generative replay is often limited to relatively simple datasets [422], [438]. An alternative is to convert the target of generative replay from data level to feature level, which can largely reduce the complexity of conditional generation and more adequately exploit semantic information. For example, GFR [273] trains conditional GANs to replay generated features after the feature extractor. BI-R [422] incorporates context-modulated feedback connections in a standard VAE to replay internal representations.

In fact, maintaining feature-level rather than data-level distributions enjoys numerous benefits in terms of efficiency and privacy. We refer to this sub-direction as **feature replay**. However, a central challenge is the *representation shift* caused by sequentially updating the feature extractor, which reflects the feature-level catastrophic forgetting. To address this issue, GFR [273], FA [181] and DSR [525] perform feature distillation between the old and new models. IL2M [31] and SNCL [139] recover statistics of feature representations (e.g., mean and covariance) on the basis of experience replay. RER [419] explicitly estimates the representation shift to update the preserved old features. REMIND [156] and ACAE-REMIND [437] instead fix the early layers of the feature extractor and reconstruct the intermediate representations to update the latter layers. FeTrIL [341] employs a fixed feature extractor learned from the initial task and replays generated features afterwards.

For continual learning from scratch, the required changes in representation are often dramatic, while stabilizing the feature extractor may interfere with accommodating new representations. In contrast, the use of strong pre-training can provide robust representations that are generalizable to downstream tasks and remain relatively stable in continual learning. An empirical study [321] has systematically investigated feature replay for continual learning with large-scale pre-training. A more in-depth discussion of this topic is presented in Sec. 4.4.

4.3 Optimization-Based Approach

Continual learning can be achieved by not only adding additional terms to the loss function (e.g., regularization and replay), but also explicitly designing and manipulating the optimization programs.

A typical idea is to perform **gradient projection**. Some replay-based approaches such as GEM [281], A-GEM [66], LOGD [411] and MER [369] constrain parameter updates to align with the direction of experience replay, corresponding to preserving the previous input space and gradient space with some old training samples. In contrast to saving old training samples, OWM [496] and AOP [146] modify parameter updates to the orthogonal direction of the previous input space. OGD [115] preserves the old gradient directions and rectifies the current gradient directions orthogonal to them. Orthog-Subspace [65] performs continual learning with orthogonal low-rank vector subspaces and keeps the gradients of different tasks orthogonal to each other. GPM [380] maintains the gradient subspace important to the old tasks (i.e., the bases of core gradient space) for orthogonal projection in updating parameters. CGP [68] calculates such gradient subspace from individual classes to additionally mitigate inter-class interference. FS-DGPM [90] dynamically releases unimportant bases of GPM [380] to improve learning plasticity and encourages the convergence to a flat loss landscape. CUBER [261] selectively projects gradients to update the knowledge of old tasks that are positively related to the current task. TRGP [260] defines the "trust region" through the norm of gradient projection onto the subspace of previous inputs, so as to selectively reuse the frozen weights of old tasks. Adam-NSCL [448] instead projects candidate parameter updates into the current null space approximated by the uncentered feature covariance of the old tasks, while AdNS [227] considers the shared part of the previous and current null spaces. NCL [202] unifies Bayesian weight regularization and gradient projection, encouraging parameter updates in the null space of the old tasks while converging to a maximum of the Bayesian approximation posterior. Under the upper bound of the quadratic penalty in Bayesian weight regularization, RGO [266] modifies gradient directions with a recursive optimization procedure to obtain the optimal solution. Therefore, as regularization and replay are ultimately achieved by rectifying the current gradient directions, gradient projection corresponds to a similar modification but explicitly in parameter updates.

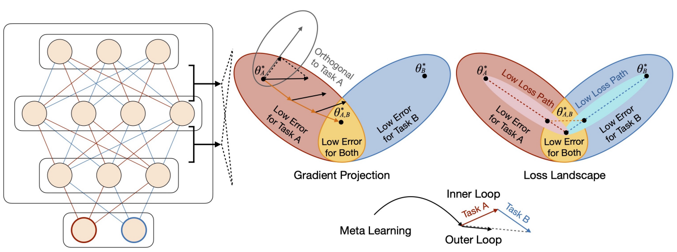


Fig. 6. Optimization-based approach. This direction is characterized by explicit design and manipulation of the optimization programs, such as gradient projection with reference to the gradient space or input space of the old tasks (adapted from [115]), meta-learning of sequentially arrived tasks within the inner loop, and exploitation of mode connectivity and flat minima in loss landscape (adapted from [258], [312]). θ_A^*, θ_B^* and $\theta_{A,B}^*$ are desirable solutions for task A, task B and both of them, respectively.

Another attractive idea is **meta-learning** or learning-to-learn for continual learning, which attempts to obtain a data-driven inductive bias for various scenarios, rather than designing it manually [150]. OML [186] provides a meta-training strategy that performs online updates on the sequentially arrived inputs and minimizes their interference, which can naturally obtain sparse representations suitable for continual learning. ANML [30] extends this idea by meta-learning of a context-dependent gating function to selectively activate neurons with respect to incremental tasks. AIM [238] learns a mixture of experts to make predictions with the representations of OML [186] or ANML [30], further sparsifying the representations at the level of architecture. Meanwhile, meta-learning can work with experience replay to better utilize both the old and new training samples. For example, MER [369] aligns their gradient directions, while iTAML [357] applies a meta-updating rule to keep them in balance with each other. With the help of experience replay, La-MAML [148] optimizes the OML [186] objective in an online fashion with an adaptively modulated learning rate. OSAKA [49] proposes a hybrid objective of knowledge accumulation and fast adaptation, which can be resolved by obtaining a good initialization with meta-training and then incorporating knowledge of incremental tasks into the initialization. Meta-learning can also be used to optimize specialized architectures. MERLIN [223] consolidates a meta-distribution of model parameters given the representations of each task, which allows to sample task-specific models and ensemble them for inference. Similarly, PR [161] adopts a Bayesian strategy to learn task-specific posteriors with a shared meta-model. MARK [176] maintains a set of shared weights that are incrementally updated with meta-learning and selectively masked to resolve specific tasks. ARI [446] combines adversarial attacks with experience replay to obtain task-specific models, which are then fused together through meta-training.

Besides, some other works refine the optimization process from a **loss landscape** perspective. For example, rather than dedicating an algorithm, Stable-SGD [313] enables SGD to find a flat local minima by adapting the factors in training regime, such as dropout, learning rate decay and batch size. MC-SGD [312] empirically demonstrates that the local minima obtained by multi-task learning (i.e., joint training of all incremental tasks) and continual learning can be connected by a linear path of low error, and applies experience replay to find a better solution along it. Linear Connector [258] adopts Adam-NSCL [448] and feature distillation to obtain respective solutions of the old and new tasks connected by a linear path of low error, followed by linear averaging. Further, un-/self-supervised learning (than traditional supervised learning) [127], [168], [286] and large-scale pre-training (than random initialization) [84], [168], [300], [360], [467] have been shown to suffer from less catastrophic forgetting. Empirically, both can be attributed to obtaining a more robust (e.g., orthogonal, sparse and uniformly-scattered) representation [168], [286], [360], [389], and converging to a wider loss basin [152], [168], [286], [300], [317], suggesting a potential link among the sensitivity of representations, parameters and task-specific errors. Many efforts seek to leverage these advantages in continual learning, as we discuss next.

4.4 Representation-Based Approach

We group the approaches that create and exploit the strengths of representations for continual learning into this category. In addition to an earlier work that acquires sparse representations from meta-training [186], recent work has attempted to incorporate the advantages of self-supervised learning (SSL) [127], [286], [343] and large-scale pre-training [300], [389], [467] to improve the representations in initialization and in continual learning. Note that these two strategies are closely related, since the pre-training data is usually of a huge amount and without explicit labels, while the performance of SSL itself is mainly evaluated by fine-tuning on (a sequence of) downstream tasks. Below, we will discuss representative sub-directions.

The first is to implement **self-supervised learning** (basically with contrastive loss) for continual learning. Observing that self-supervised representations are more robust to catastrophic forgetting, LUMP [286] acquires further improvements by interpolating between instances of the old and new tasks. MinRed [349] further promotes the diversity of experience replay by de-correlating the stored old training samples. CaSSLe [118] converts the self-supervised loss to a distillation strategy by mapping the current state of a representation to its previous state. Co2L [57] adopts a supervised contrastive loss to learn each task and a self-supervised loss to distill knowledge between the old and new models. DualNet [343] trains a fast learner with supervised loss and a slow learner with self-supervised loss, with the latter helping the former to acquire generalizable representations.

The second is to use **pre-training for downstream continual learning**. Several empirical studies suggest that downstream continual learning clearly benefits from the use of pre-training, which brings not only strong knowledge transfer but also robustness to catastrophic forgetting [127], [300], [321], [360]. In particular, the benefits for downstream continual learning tend to be more apparent when pre-training with larger data size [321], [360], larger model size [360] and contrastive loss [84], [127]. However, a critical challenge is that the pre-trained knowledge needs to be adaptively leveraged for the current task while maintaining generalizability to future tasks. There are various strategies for this problem, depending on whether the pre-trained backbone is fixed or not.

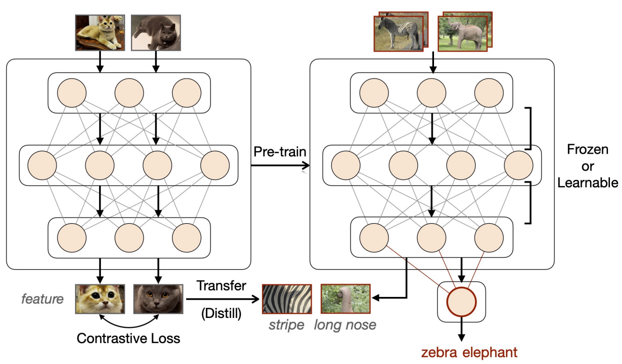


Fig. 7. Representation-based approach. This direction is characterized by creating and leveraging the strengths of representations for continual learning, such as by using self-supervised learning and pre-training. In particular, both upstream pre-training and downstream fine-tuning require continual learning, while the pre-trained representations are optionally fixed for learning specific downstream tasks.

As for adapting a *fixed* backbone, Side-Tuning [506] and DLCFT [394] train a lightweight network in parallel with the backbone and fuse their outputs linearly. TwF [42] also trains a sibling network, but distills knowledge from the backbone in a layer-wise manner. GAN-Memory [81] takes advantage of FiLM [339] and AdaFM [515] to learn task-specific parameters for each layer of a pre-trained generative model, while ADA [113] employs Adapters [166] with knowledge distillation to adjust a pre-trained transformer. Recent prompt-based approaches instruct the representations of a pre-trained transformer with a few prompt parameters. Such methods typically involve construction of task-adaptive prompts and inference of appropriate prompts for testing, by exploring prompt architectures to accommodate task-sharing and task-specific knowledge. Representative strategies include selecting the most relevant prompts from a prompt pool (L2P [458]), performing a weighted summation of the prompt pool with attention factors (CODA-Prompt [400]), using explicitly task-sharing and task-specific prompts (DualPrompt [457]) or only task-specific prompts (S-Prompts [450], HiDe-Prompt [439]), progressive expansion of task-specific prompts (Progressive Prompts [364]), etc. Besides, by saving prototypes, appending a nearest class mean (NCM) classifier to the backbone has proved to be a strong baseline [185], [332], which can be further enhanced by transfer learning techniques such as the FiLM adapter [327]. As for optimizing an *updatable* backbone, F2M [387] searches for flat local minima in the pre-training stage, and then learns incremental tasks within the flat region. CwD [389] regularizes the initial-phase representations to be uniformly scattered, which can empirically mimic the representations of joint training. SAM [122], [300] encourages finding a wide basin in downstream continual learning by optimizing the flatness metric. SLCA [503] observes that slowly fine-tuning the backbone of a pre-trained transformer can achieve outstanding performance in continual learning, and further preserves prototype statistics to rectify the output layer.

The third is **continual pre-training** (CPT) or continual meta-training. As the huge amount of data required for pre-training is typically collected in an incremental manner, performing upstream continual learning to improve downstream performance is particularly important. For example, an empirical study finds that self-supervised pre-training is more effective than supervised protocols for continual learning of vision-language models [82], consistent with the results for only visual tasks [168]. Since texts are generally more efficient than images, IncCLIP [478] replays generated hard negative texts conditioned on images and performs multi-modal knowledge distillation for updating CLIP [355]. For CPT of language models, ECONET [151] designs a self-supervised framework with generative replay. Meanwhile, continual meta-training needs to address a similar issue that the pre-trained knowledge of base classes is incrementally enriched and adapted. IDA [268] imposes discriminants of the old and new models to be aligned relative to the old centers, and otherwise leaves the embedding free to accommodate new tasks. ORDER [456] employs unlabeled OOD data with experience replay and feature replay to cope with large inter-task differences.

4.5 Architecture-Based Approach

The above strategies basically focus on learning all incremental tasks with a shared set of parameters (i.e., a single model as well as one parameter space), which is a major cause of the inter-task interference. In contrast, constructing task-specific parameters can explicitly resolve this problem. Previous work generally separates this category into *parameter isolation* and *dynamic architecture*, depending on whether the network architecture is fixed or not. Here, we instead focus on the way of implementing task-specific parameters, extending the above concepts to parameter allocation, model decomposition and modular network (see Fig. 8).

Parameter allocation features an isolated parameter subspace dedicated to each task throughout the network, where the architecture can be fixed or dynamic in size. Within a *fixed* network architecture, Piggyback [290], HAT [384], SupSup [463], MEAT [477], WSN [199] and H^2 [190] explicitly optimize a binary mask to select dedicated neurons or parameters for each task, with the masked regions of the old tasks (almost) frozen to prevent catastrophic forgetting.

PackNet [291], UCL [9], CLNP [138], AGS-CL [196] and NISPA [149] explicitly identify the important neurons or parameters for the current task and then release the unimportant parts to the following tasks, which can be achieved by iterative pruning [291], activation value [138], [149], [196], uncertainty estimation [9], etc. Since the network capacity is limited, "free" parameters tend to saturate as more incremental tasks are introduced. Therefore, these methods typically require sparsity constraints on parameter usage and selective reuse of the frozen old parameters, which might affect the learning of each task. To alleviate this dilemma, the network architecture can be *dynamically expanded* if its capacity is not sufficient to learn a new task well, such as DEN [492], CPG [175] and DGMa/DGMw [322]. The dynamic architecture can also be explicitly optimized to improve parameter efficiency and knowledge transfer, such as by reinforcement learning (RCL [475], BNS [352]), architecture search (LtG [254], BNS [352]), variational Bayes (BSA [232]), etc. As the network expansion should be much slower than the task increase to ensure scalability, constraints on sparsity and reusability remain important.

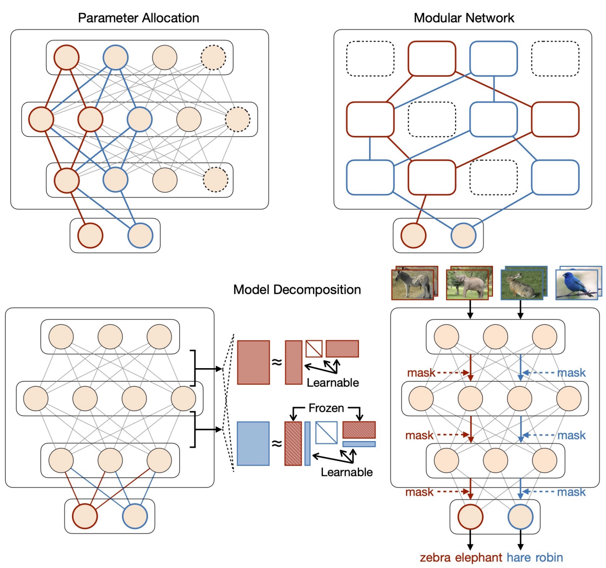


Fig. 8. Architecture-based approach. This direction is characterized by constructing task-specific/adaptive parameters with a properly-designed architecture, such as assigning dedicated parameters to each task (parameter allocation), constructing task-adaptive sub-modules or sub-networks (modular network), and decomposing the model into task-sharing and task-specific components (model decomposition). Here we exhibit two types of model decomposition, corresponding to parameters (low-rank factorization, adapted from [178]) and representations (masking of intermediate features).

Model decomposition separates a model explicitly into task-sharing and task-specific components, where the task-specific components are often expandable. For a regular network, the task-specific components could be parallel branches (ACL [109], ReduNet [470], EPIE-Net [106]), adaptive layers (GVCL [280], DyTox [103]), masks or mask generators for intermediate features (CCLL [398], CCG [1], MARK [176]). Note that the feature masks for model decomposition do not operate in parameter space and are not binary for each task, thus fundamentally different from the binary masks for parameter allocation. Besides, the network parameters themselves can be decomposed into task-sharing and task-specific elements, such as by additive decomposition (APD [490]), singular value decomposition (RCM [198]), filter atom decomposition (FAS [304]) and low-rank factorization (IBP-WF [299], IRU [178]). As the number of task-specific components usually grows linearly with incremental tasks, their resource efficiency determines the scalability of this sub-direction.

Modular network leverages parallel sub-networks or sub-modules to learn incremental tasks in a differentiated manner, without pre-defined task-sharing or task-specific components. As an early work, Progressive Networks [379] introduces an identical sub-network for each task and allows knowledge transfer from other sub-networks via adaptor connections. Expert Gate [14] and a subsequent work [80] employ a mixture of experts [184] to learn incremental tasks, expanding one expert as each task is introduced. PathNet [117] and RPSNet [356] pre-allocate multiple parallel networks to construct a few candidate paths, i.e., layer-wise compositions of network modules, and select the best path for each task. MNTDP [427] and LMC [323] attempt to find the optimal layout from previous sub-modules and potentially new sub-modules. Similar to parameter allocation, these efforts are intentional to construct task-specific models, while the combination of sub-networks or sub-modules allows explicit reuse of knowledge. In addition, the sub-networks can be encouraged to learn incremental tasks in parallel. Model Zoo [361] expands a sub-network to learn each new task with experience replay of the old tasks, and ensembles all sub-networks for prediction. CoSCL [443] and CAF [442] ensembles multiple continual learning models and modulates the predictive similarity between them, proving to be effective in resolving the discrepancy of task distribution and improving the flatness of loss landscape.

In a broader sense, sampling parameters from task-conditioned parameter distributions (MERLIN [223], PR [161], PGMA [170], HNET [431]), as well as stabilizing important parameters with weight regularization, can be seen as a form of deriving task-specific/adaptive parameters. In contrast to other directions, most architecture-based approaches amount to de-correlating incremental tasks in network parameters, which can almost avoid catastrophic forgetting but affect scalability and inter-task generalizability. In particular, task identities are often required to determine which set of parameters to use, thus greatly constraining their applications. To overcome this limitation, task identities can be inferred from the responses (e.g., predictive uncertainty) of each task-specific model [14], [161], [218]. The function of task-identity prediction can also be learned explicitly from incremental tasks, using other continual learning strategies to mitigate catastrophic forgetting [1], [80], [109], [161], [190], [223], [299].

Besides, the design and choice of a **basic architecture** can largely impact the continual learning performance. For example, wider networks tend to be more robust to catastrophic forgetting due to more orthogonality and sparsity in gradient directions, as well as a lazier training regime [309], [310]. Batch Normalization (BN) layers [180] tend to introduce biased moments towards the current task, resulting in sub-optimal performance of the previous tasks [59], [284], [345]. Dropout [403] in a stable network behaves like a gating mechanism to create task-specific pathways and thus can mitigate catastrophic forgetting [311].

5 Scenario Complexity in Application: The Case of Visual Classification

Due to the complexity of the real world, practical applications present a variety of specialized challenges. Here we categorize these challenges into **scenario complexity** and **task specificity**, with an extensive analysis of how continual learning methods are adapted to them. In this section, we demonstrate the challenges of scenario complexity, using visual classification as an example.

5.1 Task-Agnostic Inference

Continual learning usually has *Task-Incremental Learning* (TIL) as a basic setup, i.e., task identities are provided in both training and testing. In contrast, task-agnostic inference that avoids the use of task identities for testing tends to be more natural but more challenging in practical applications, which is known as *Class-Incremental Learning* (CIL) for classification tasks. For example, let's consider two binary classification tasks: (1) "zebra" and "elephant"; and (2) "hare" and "robin". After learning these two tasks, TIL needs to know which task it is and then classify the two classes accordingly, while CIL directly classifies the four classes at the same time. Therefore, CIL has received increasingly more attention and become almost the most representative scenario for continual learning.

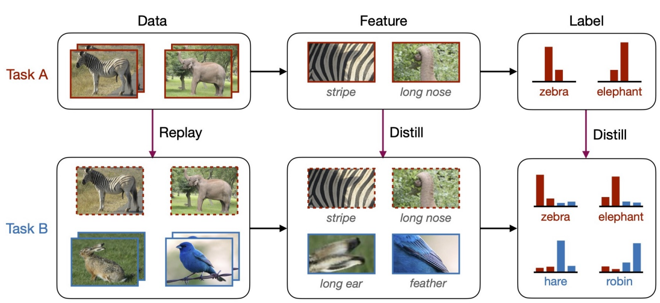


Fig. 9. Representative strategies for class-incremental learning. Catastrophic forgetting can be mitigated with respect to data space (experience replay), feature space (knowledge distillation) and label space (knowledge distillation). This figure is adapted from [172].

The CIL problem can be disentangled into *within-task prediction* and *task-identity prediction*, where the latter is a particular challenge posed by task-agnostic inference and has been shown to be closely related to the OOD detection [218]. To overcome catastrophic forgetting in CIL, numerous efforts attempt to impose the behaviors of the previous model on the current model, in terms of data, feature, and label spaces (see Fig. 9). Since the incremental classes are disjoint, the new training samples are usually OOD from the old ones. In this regard, replay of the old training samples is able to impose an end-to-end effect, while feature and label distillations of the new training samples are limited to the output side with a biased distribution [172]. Therefore, many state-of-the-art methods are built on the framework of *experience replay* and then incorporate *knowledge distillation*, as we have discussed in Sec. 4.2. In Table 2, we summarize these CIL methods based on their main focus.

TABLE 2
Summary of representative CIL methods with the use of experience replay. These methods further improve the memory buffer, feature distillation or label distillation to achieve better performance, corresponding to data, feature and label spaces, respectively.

Main Focus	Representative Method
Data Space	iCaRL [365], GSS [16], Mnemonics [276], TPCIL [413], GDumb [348], DER++ [46], RMM [275], HAL [64], MRDC [444], CSI [218], X-DER [41]
Feature Space	LUCIR [165], PODNet [102], TPCIL [413], PCL [171], AANets [274], DER [479], DDE [172], GeoDL [396], PASS [523], Co2L [57], AFC [200], SP-CIL [467], ELI [193], CwD [389], CSCCT [21], FOSTER [435], FASP [304], CLS-ER [17]
Label Space	LwF [255], iCaRL [365], GEM [281], A-GEM [66], EEIL [53], BiC [468], WA [512], DER++ [46], ScaIL [32], S&B [219], SS-IL [10], Coil [519]

To avoid the additional resource overhead and potential privacy issues of retaining old training samples, many efforts attempt to perform CIL without experience replay, i.e., *Data-Free* CIL. An intriguing idea is to replay synthetic data produced by inverting a frozen copy of the old classification model, such as DeepInversion [487], ABD [399], RRL [129] and CF-IL [347], which usually further incorporate knowledge distillation to compensate the lost information in model inversion. Other methods exploit the class-wise statistics of feature representations to obtain a balanced classifier, such as by imposing the representations to be transferable and invariant (SPB [465], IL2A [522], SSRE [525], FeTrIL [341]), or compensating explicitly the representation shifts (SDC [495], RER [419]). Besides, the use of adequate pre-training provides well-distributed representations, which allows strong CIL performance without the need of experience replay. Representative strategies include prompting [400], [439], [457], [458] or carefully fine-tuning [300], [387], [503] the backbone, saving prototypes to rectify the predictions [185], [332], [503], incorporating transfer learning techniques [42], [113], [327], etc., as discussed in Sec. 4.4.

For continual learning (especially CIL) of visual classification tasks, the current state-of-the-art methods mainly focus on image classification with attention to relatively complex and large-scale datasets, such as ILSVRC2012 [228] and its derivatives. There are also many benchmarks for video classification [279], [330], [372], [429], varying in size and purpose.

5.2 Scarcity of Labeled Data

Most of the current continual learning settings assume that incremental tasks have sufficiently large amounts of labeled data, which is often expensive and difficult to obtain in practical applications. For this reason, there is a growing body of work focusing on the scarcity of labeled data in continual learning.

A representative scenario is *Few-Shot Continual Learning* (FSCL) or further specified as *Few-Shot* CIL (FSCIL) [414], where the model first learns some base classes for initialization with a large number of training samples, and then learns a sequence of novel classes with only a small number of training samples. The extremely limited training samples exacerbate the overfitting of previously-learned representations to subsequent tasks, which can be alleviated by recent work such as preserving the topology of representations (TOPIC [414]), constructing an exemplar relation graph for knowledge distillation (ERL [99]), selectively updating only unimportant parameters (FSLL [295]) or stabilizing important parameters (LCwoF [230]), cooperating fast-slow weights (MgSvF [513]), updating parameters within the flat region of loss landscape (F2M [387]), meta-learning of a good initialization (MetaFSCIL [74]), and generative replay of old data distributions (ERDFR [265]).

There are many other efforts keeping the backbone fixed in subsequent continual learning, so as to decouple the learning of *representation* and *classifier*. Under this framework, representative strategies can be conceptually separated into two aspects. The first is to obtain compatible and extensible representations from massive base classes, such as by enforcing the representations compatible with simulated incremental tasks (SPPR [524], LIMIT [518]), reserving the feature space with virtual prototypes for future classes (Fact [517]), using angular penalty loss with data augmentation (ALICE [336]), using self-supervised learning (S3C [197]), providing extra constraints from margin-based representations (CLOM [527]), etc. The second is to obtain an adaptive classifier from a sequence of novel classes, such as by evolving the classifier weights with a graph attention network (CEC [501]), developing a tree-based hierarchical classifier with Gaussian processes (GP-Tree [6]), performing hyperdimensional computing (C-FSCIL [162]), and sampling stochastic classifiers from a weight distribution (S3C [197]). Besides, auxiliary information such as semantic word vectors (SKD [72], Mixture of Subspaces [73], Subspace Reg [11], FSIL-GAN [8]) and sketch exemplars (DIY [37]) can be incorporated to enrich the limited training samples.

In addition to a few labeled data, there is usually a large amount of unlabeled data available and collected over time. The first practical setting is called *Semi-Supervised Continual Learning* (SSCL) [440], which considers incremental data as partially labeled. As an initial attempt, ORDisCo [440] learns a semi-supervised classification model together with a conditional GANs for generative replay, and regularizes discriminator consistency to mitigate catastrophic forgetting. Subsequent work includes training an adversarial autoencoder to reconstruct images (AAE [237]), imposing predictive consistency among augmented and interpolated data (CCIC [43]), leveraging the nearest-neighbor classifier to distill class-instance relationships (NNCSL [201]), etc. The second scenario assumes that there is an external unlabeled dataset to facilitate supervised continual learning, e.g., by knowledge distillation (GD [241]) and data augmentation (L2I [412]). The third scenario is to learn representations from incremental unlabeled data [82], [168], i.e., *Unsupervised Continual Learning* (UCL), which becomes an increasingly important topic for updating pre-trained knowledge in foundation models.

5.3 Generic Learning Paradigm

Potential challenges of the learning paradigm can be summarized in a broad concept called *General Continual Learning* (GCL) [15], [46], [86], where the model observes incremental data in an online fashion without explicit task boundaries. Correspondingly, GCL consists of two interconnected settings: *Task-Free Continual Learning* (TFCL) [15], where the task identities are not accessible in either training or testing; and *Online Continual Learning* (OCL) [16], where the training samples are observed in an one-pass data stream. Since TFCL usually accesses only a small batch of training samples at a time for gradual changes in task distributions, while OCL usually requires only the data label rather than the task identity for each training sample, many representative methods for TFCL and OCL are indeed compatible (see Table 3 for their target scenarios).

TABLE 3
Summary of TFCL, OCL and GCL (both TFCL and OCL) methods. Of note, some methods are not designed for specific scenarios, but are used as strong baselines in subsequent work.

Scenario	Representative Method
TFCL	CN-DPM [244], [188], VariGrow [18], ODDL [484]
OCL	InstAParam [69], CBRS [76], ASER [391], InfoRS [408], DVC [145], CTN [344], NCCL [486], ILOS [159], SCR [289], CVT [453], OCM [147], ER-ACE [47], RAR [510], PoLRS [50], AOP [146], CV [160]
GCL	GSS [16], MIR [13], CLIB [226], GMED [192], DRO [455], GDumb [348], DER [46], GEM [281], A-GEM [66], DSDM [346], CoPE [87], BLD [119]

Specifically, some of them attempt to learn specialized parameters in a growing architecture. For TFCL, CN-DPM [244] adopts Dirichlet process mixture models to construct a growing number of neural experts, while a concurrent work [188] derives such mixture models from a probabilistic meta-learner. VariGrow [18] employs an energy-based novelty score to decide whether to extend a new expert or update an old one. ODDL [484] estimates the discrepancy between the current memory buffer and the previously-learned knowledge as an expansion signal. For OCL, InstAParam [69] selects and consolidates appropriate network paths for individual training samples.

On the other hand, many TSCL methods and most OCL methods are built on experience replay, focusing on construction, management and exploitation of a memory buffer. As the training samples arrive in small batches, the information of task boundaries is less effective, and reservoir sampling usually serves as an effective baseline strategy for sample selection. More advanced strategies prioritize the replay of those training samples that are informative (InfoRS [408]), diversified in parameter gradients (GSS [16]) or previously-learned knowledge (ODDL [484]), balanced in class labels (CBRS [76], GDumb [348], CoPE [87]), and beneficial to latent decision boundaries (ASER [391]). Meanwhile, the memory buffer can be dynamically managed, such as by removing less important training samples (CLIB [226]), editing the old training samples (GMED [192]) or their distributions (DRO [455]) to be more likely forgotten, and retrieving the old training samples that are susceptible to interference (MIR [13], DVC [145]). To adaptively exploit the memory buffer, representative strategies include using additional parameters to construct task-specific features (CTN [344]) or calibrate the network (NCCL [486]), performing knowledge distillation (ILOS [159], DER [46]), evolving prototypes in feature space (CoPE [87]), improving representations with contrastive learning (SCR [289], DVC [145], CVT [453]), maximizing mutual information between input data and their label predictions (OCM [147]), using asymmetric cross-entropy (ER-ACE [47]) or constrained gradient directions (GEM [281], A-GEM [66]) of the old and new training samples, repeated rehearsal with data augmentation (RAR [510]), properly adjusting the learning rate (PoLRS [50], CLIB [226]), training from scratch with the memory buffer (GDumb [348]), etc.

In addition, there are some OCL methods without using experience replay, such as by distilling knowledge from the current batch of training samples (BLD [119]), and using a pre-trained backbone for orthogonal gradient projection (AOP [146]) or feature replay (CV [160]).

6 Task Specificity in Application

In this section, we describe the challenges of task specificity in various realistic applications of continual learning.

6.1 Object Detection

Incremental Object Detection (IOD) is a typical extension of continual learning for object detection, where the training samples annotated with different classes are introduced in sequence, and the model needs to correctly locate and identify the object instances belonging to the previously-learned classes. Unlike visual classification with only one object instance appearing in each training sample, object detection usually has multiple object instances belonging to the old and new classes appearing together. Such *co-occurrence* poses an additional challenge for IOD, where the object instances of old classes are annotated as the "background" when learning new classes, thus exacerbating catastrophic forgetting. On the other hand, this makes knowledge distillation a naturally powerful strategy for IOD, since the object instances of old classes can be obtained from new training samples to constrain the differences in responses between the old and new models.

As an early work, ILOD [393] distills the responses for old classes to prevent catastrophic forgetting on Fast R-CNN [135], followed by RKT [358] that further distills the relation of co-occurrence in selected proposals. The idea of knowledge distillation is then introduced to other object detectors, such as SID [335] on CenterNet [521], RILOD [250] on RetinaNet [262], ERD [116] on GFLV1 [253], CIFRCN [153], Faster ILOD [334], DMC [505], BNC [98] and IOD-ML [224] on Faster R-CNN [367], etc. Some methods exploit the unlabeled in-the-wild data to distill the old and new models into a shared model, in order to bridge potential non co-occurrence (BNC [98]) and to achieve a better stability-plasticity trade-off (DMC [505]). To further mitigate the negative effects of knowledge distillation on learning plasticity, IOD-ML [224] adopts meta-learning to reshape parameter gradients into a balanced direction between the old and new classes.

IOD is not only applicable for 2D images, but also for 3D images [516] and videos [436]. Besides, there are many other related settings, such as incremental few-shot detection [340], where a pre-trained object detector registers new classes with only a few annotated data; and open world object detection [194], where the object detector needs to identify potential object instances of unknown classes and register them after receiving corresponding annotations.

6.2 Semantic Segmentation

Continual Semantic Segmentation (CSS) aims at pixel-wise prediction of classes and learning new classes in addition to the old ones. Similar to IOD, the object instances of old and new classes can appear together. Some early efforts utilize full annotations of both the old and new classes in continual learning [305], [307]. However, due to the significant expense and time cost of re-annotating the old classes, more attention has been focused on using annotations of only the new classes, which leads to the old classes being treated as the background (known as the *background shift*) and thus exacerbates catastrophic forgetting.

A common strategy is to distill knowledge adaptively from the old model, which can faithfully distinguish unannotated old classes from the background. For example, MiB [56] calibrates regular cross-entropy (CE) and knowledge distillation (KD) losses of the background pixels with predictions from the old model. ALIFE [319] further improves the calibrated CE and KD with logit regularization, and fine-tunes the classifier with feature replay. RCIL [502] reparameterizes the network into two parallel branches, where the old branch is frozen for KD between intermediate layers. SDR [306] and UCD [481] introduce contrastive learning into distillation of latent representations, where pixels of the same class are clustered and pixels of different classes are separated. PLOP [101], RECALL [292], SSUL [61], EM [480], Self-Training [494], UCD [481] and WILSON [55] explicitly use the old model to generate pseudo-labels of the old classes. Auxiliary data resources such as a web crawler (RECALL-Web [292]), a pre-trained generative model (RECALL-GAN [292]), large amounts of unlabeled data (Self-Training [494]), and a few old training samples (EM [480], SSUL [61], ILLR [123]) have been exploited to facilitate KD and prevent catastrophic forgetting. Besides, saliency maps are commonly used to locate unannotated objects in CSS, in response to weak supervision of only image-level annotations (WILSON [55], ILLR [123]), as well as defining unknown classes within the background to benefit learning plasticity (SSUL [61]).

In addition to the regular CSS, other relevant settings include unsupervised domain adaptation for CSS [404], where a pre-trained backbone adapts to the unannotated data of the target domain without using the annotated data of the source domain; incremental few-shot semantic segmentation [388], which performs CSS with only a few annotated data; incremental instance segmentation [144], which incrementally learns new classes but requires individual segmentation for each instance; as well as incremental few-shot instance segmentation [128].

6.3 Conditional Generation

Continual Learning for Conditional Generation (CLCG) is closely related to generative replay, which can mitigate catastrophic forgetting of the currently-trained generative model and/or discriminative model through recovering the previously-learned data distributions. This usually requires saving a frozen copy of the old generative model for conditional generation during a new training process. Since the sampling distributions are often provided in both training and testing, CLCG is similar to the setting of TIL, without the challenge of task-identity prediction.

As a generic framework, DGR [392] performs generative replay for continual learning of both discriminative and generative models. Many efforts focus on CLCG with GANs, such as by weight regularization (ORDisCo [440], [383]), generative replay (MeRGANs [464], L-VAEGAN [483]), knowledge distillation (LifelongGAN [500], Hyper-LifelongGAN [499]), parameter allocation (Piggyback-GAN [498], DGMa/DGMw [322], TMNs [438]), model decomposition (Hyper-LifelongGAN [499], FILIT [70], GAN-Memory [81]), etc. Other work performs CLCG with VAE, which often relies on a multi-head architecture with encoder expansion, as well as weight regularization (VCL [318]) or generative replay (CURL [363], VASE [5]) to overcome catastrophic forgetting. BooVAE [112] instead adopts a static architecture through incorporating each task into an additive aggregated posterior and using it as the prior for the next task.

6.4 Reinforcement Learning

Continual Reinforcement Learning (CRL) is required to address dynamic data distributions between and within tasks (corresponding to the settings of TIL and TFCL, respectively), because of the over-time interactions of the states, actions and environments. Various types of continual learning strategies have been shown to work not only for visual classification but also for reinforcement learning (RL), such as EWC [222], MAS [12], VCL [318], P&C [382], AFEC [441], UCL [9], AGS-CL [196], CPR [60], Progressive Networks [379], PackNet [291], ER, A-GEM [66], etc.

With respect to specialized strategies, bio-inspired weight regularization (Benna-Fusi [204]) and function regularization (PC [205]) are proposed to mitigate within-task catastrophic forgetting. To cope with multiple tasks, OWL [213] implements EWC [222] in a shared feature extractor of separate output heads, and adopts multi-armed bandit for task-identity prediction. An empirical study [182] evaluates different principles of selecting old experiences for replay, with matching training distributions usually performing best. CLEAR [373] mixes off-policy learning from old experiences and on-policy learning from novel experiences, with behavioral cloning to further promote stability. MTR [203] employs a cascade of interacting sub-buffers to accumulate experiences at different timescales. LPG-FTW [301] instead factorizes a policy gradient model into task-sharing and task-specific parameters. ClonEx-SAC [460] investigates the effects of actor, critic, exploration and experience replay on knowledge transfer and provides a set of general recommendations. COMPS [35] considers continual meta-training of RL, using behavioral cloning to quickly learn new tasks from the knowledge of previous incremental tasks.

A distinctive feature of CRL is its high diversity of applications and benchmarks, including continuous control [205], [283], [301], [461], maze navigation [278], [379], video games (Atari [9], [60], [196], [222], [379], [441], StarCraft [381], Minecraft [415], Hanabi [316]), etc., varying widely in task type and computation overhead [461]. A systematic comparison of CRL methods in various contexts is a promising future work.

6.5 Natural Language Processing

Continual learning in *Natural Language Processing* (NLP) has received increasing attentions in recent years, which shares many typical settings and representative methods proposed in visual domains. Specifically, a range of continual learning scenarios have been considered for NLP, including DIL, TIL, CIL, OCL and CPT [209]. Accordingly, many representative methods are adapted to these scenarios and proven to be effective, such as weight regularization (RMR-DSE [249], SRC [270]), knowledge distillation (ExtendNER [315], CFID [251], CID [269], PAGeR [425], LFPT5 [350], DnR [406], CL-NMT [51], COKD [385]), experience replay (CFID [251], CID [269], ELLE [353], IDBR [173], MBPA++ [88], Meta-MBPA++ [454], EMAR [476], DnR [406], ARPER [302], Total Recall [256]), generative replay (PAGeR [425], LAMOL [405], ACM [511], NER [447]), parameter allocation (TPEM [133]), modular network (ProgModel [386]), meta-learning (Meta-MBPA++ [454], MeLL [433], CML [466]), etc.

On the other hand, continual learning in NLP is characterized by the extensive use of pre-training in transformer architectures, which motivates the implementations of parameter-efficient fine-tuning techniques. The pre-trained transformer can effectively accommodate successive arrivals by learning only a few task-specific parameters at a time, including *adaptor-tuning* [166] with inserted fully-connected layers (CPT [208], CLIF [191], AdapterCL [287], ACM [511], ADA [113]) and *prompt-tuning* [247] with trainable prompt tokens (C-PT [526], LFPT5 [350], EMP [267]). Task specificity can also be acquired by *instruction* [111] in continual learning, which adds a short text to describe the core concept of each task (PAGeR [425], ConTinTin [488], ENTAILMENT [471]). Due to the great success of pre-trained foundation models, these techniques are increasingly being used and further facilitate continual learning in visual domains [113], [439], [450], [451], [453], [457], [458].

Besides, the applications of NLP in conjunction with continual learning are highly diversified, providing unique opportunities for subsequent research. Representative tasks that have been explored include dialogue system [132], [133], [264], [287], [302], [449], text classification [88], [173], [331], [471], sentence generation [249], [302], [511], relation learning [351], [366], [466], [476], neural machine translation [51], [130], [143], [385], named entity recognition [315], [350], [447], etc. Some other work considers multi-modality of vision and language for continual pre-training [82], [478] or downstream tasks [89], [142], [402]. We refer readers to their original papers for technical details.

6.6 Beyond Task Performance

Continual learning can benefit many considerations beyond task performance, such as efficiency, privacy and robustness. A major purpose of continual learning is to avoid retraining all old training samples and thus improve

resource efficiency of model updates, which is not only applicable to learning multiple incremental tasks, but also important for regular single-task training. Due to the nature of gradient-based optimization, a network tends to "forget" the observed training samples and thus requires repetitive training to capture a distribution, especially for some hard examples [20], [62], [150]. Recent work has shown that the one-pass performance of visual classification can be largely improved by experience replay of hard examples [169] or orthogonal gradient projection [308]. Similarly, resolving within-task catastrophic forgetting can facilitate reinforcement learning [204], [205] and stabilize the training of GANs [242], [416].

Meanwhile, continual learning is relevant to two important directions of **privacy protection**. The first is *Federated Learning* [252], [298], where the server and clients are not allowed to communicate with data. A typical scenario is that the server aggregates the locally trained parameters from multiple clients into a single model and then sends it back. As the incremental data collected by clients is dynamic and variable, federated learning needs to overcome catastrophic forgetting and facilitate knowledge transfer across clients, i.e., federated continual learning. To achieve these aims, FedWeIT [489] decomposes network parameters into global federated parameters and sparse task-specific parameters, with the latter selectively aggregated. FedSpeech [189] adopts a set of gradual pruning masks for parameter allocation and a set of selective masks to reuse old knowledge. GLFC [97] overcomes catastrophic forgetting through knowledge distillation of a few old training samples, and alleviates class imbalance through gradient compensation. CFeD [285] instead performs knowledge distillation with auxillary unlabeled data.

The second is *Machine Unlearning*, which aims to eliminate the influence of specific training samples when their access is lost while without affecting other knowledge. Representative methods in this direction are closely related to continual learning, such as learning separate models with subsets of training samples [44], utilizing historical parameters and gradients [469], removing old knowledge from parameters with Fisher information matrix [137], adding adaptive parameters to a pre-trained backbone [136], etc. In the context of their intersection, retaining all old knowledge for continual learning may suffer from data leakage and privacy invasion, especially regarding data-level replay. Mnemonic Code [390] embeds a class-specific code when learning each class, enabling to selectively forget them through discarding the corresponding codes. LIRF [485] designs a distillation framework to remove specific old knowledge and store it in a pruned lightweight network for selective recovery.

As a strategy for adapting to variable inputs, continual learning can assist a robust model to eliminate or resist external disturbances. For example, PIGWM [520] learns adaptive mapping functions for image de-raining and overcomes catastrophic forgetting by weight regularization. NACL [376] enables face antispoofing systems to detect potential novel attacks and remember the old ones with experience replay. To cope with noisy labels of incremental data, SPR [217] obtains self-supervised representations to purify a memory buffer and then performs supervised finetuning. In fact, **robustness** and continual learning are intrinsically linked, as they correspond to generalizability in the spatial and temporal dimensions, respectively. Many ideas for improving robustness to adversarial examples have been used to improve continual learning, such as flat minima [46], [313], model ensemble [443], Lipschitz continuity [39] and adversarial training [493]. Subsequent work could further interconnect excellent ideas from both fields, e.g., designing particular algorithms and training paradigms to actively "forget" [441] external disturbances.

7 Discussion

In this section, we present an in-depth discussion of related topics in continual learning, including our observation of current trends, cross-directional prospects and interdisciplinary connections with neuroscience.

7.1 Observation of Current Trend

As continual learning is directly affected by catastrophic forgetting, previous efforts seek to address this problem by promoting memory stability of the old knowledge. However, recent work has increasingly focused on facilitating learning plasticity and inter-task generalizability. This essentially advances the understanding of continual learning: a desirable solution requires a proper balance between the old and new tasks, with adequate generalizability to accommodate their distribution differences.

To promote learning plasticity on the basis of memory stability, emergent strategies include renormalization of the old and new task solutions [164], [258], [382], [441], balanced exploitation of the old and new training samples [10], [165], [172], [452], [468], [512], reserving space for subsequent tasks [9], [196], [517], etc. On the other hand, solution generalizability can be explicitly improved by converging to a flat loss landscape, where representative strategies include optimizing the flatness metric [90], [300], [312], [313], [387], constructing an ensemble model at either spatial scale [60], [442], [443] or temporal scale [41], [46], and obtaining well-distributed representations [57], [168], [286], [300], [343], [360]. In particular, since self-supervised and pre-trained representations are naturally more robust to catastrophic forgetting [168], [286], [300], [360], creating, improving and exploiting such representational advantages has become a promising direction.

We also observe that the applications of continual learning are becoming more diverse and widespread. In addition to various scenarios of visual classification, current extensions of continual learning have covered many other vision domains, as well as other areas such as RL, NLP and ethic considerations. We present only some representative applications in the above two sections, with other more specialized and cross-cutting scenarios to be explored, such as robotics [24], [77], [246], graph learning [220], [434], [508], bioimaging [509], etc. Notably, existing work on applications has focused on providing basic benchmarks and baseline methods. Future work could develop more specialized methods to obtain stronger performance, or evaluate the generality of current methods in different applications.

7.2 Cross-Directional Prospect

Continual learning demonstrates vigorous vitality, as most of the state-of-the-art AI models require flexible and efficient updates, and their advances have contributed to the development of continual learning. Here, we discuss some attractive intersections of continual learning with other topics of the broad AI community:

Diffusion Model [29], [94], [282], [401] is a rising state-of-the-art generative model, which constructs a Markov chain of discrete steps to progressively add random noise for the input and learns to gradually remove the noise to restore the original data distribution. This provides a new target for continual learning of generative models, and its outstanding performance in conditional generation can further facilitate the efficacy of generative replay.

Foundation Model [45], [92], [355] acquires impressive results in a variety of downstream tasks due to the effective use of large-scale pre-training. The pre-training data is usually huge in volume and collected incrementally, creating urgent demands for efficient updates. Meanwhile, increasing the scale of pre-training would facilitate knowledge transfer and mitigate catastrophic forgetting between downstream tasks [300], [360]. However, fine-tuning the foundation model tends to forget pre-trained knowledge [263], which has become a central challenge for its application and requires specialized continual learning strategies.

Transformer-Based Architecture [277], [426] has proven effective for both language and vision domains, and become the dominant choice for state-of-the-art foundation models. This requires specialized designs to overcome catastrophic forgetting [451], [453] while providing new insights for maintaining task specificity in continual learning [450], [457]. The parameter-efficient transfer learning techniques developed in NLP can serve as a good reference and are being widely adapted to continual learning.

Multi-Modality [221], [320], [355], especially for contrastive learning of vision-language pairs and language-grounded applications in vision domains, has largely advanced many directions in machine learning. As an additional source of supervision, stabilization of multi-modal information can potentially mitigate catastrophic forgetting. The use of (multi-modal) large-language models (LLMs), in particular, provides strong ability of predicting task identity in both training and testing.

Embodied AI [75], [107], [125], an emerging paradigm shift from the era of "internet AI", aims to enable AI algorithms and agents to learn through interactions with their environments rather than datasets of images, videos or texts collected primarily from the Internet. Of important, the study of general continual learning helps the embodied agents to learn from an egocentric perception similar to humans, and provides a unique opportunity for researchers to pursue the essence of lifelong learning by observing the same person in a long time span [126], [436].

7.3 Connection with Neuroscience

Inspirations from neuroscience have played an important role in the development of continual learning. As biological learning is naturally on a continual basis [150], [229], [514], its underlying mechanisms provide an excellent reference for AI models. Here, we briefly overview the neurological basis of continual learning strategies, from the levels of synaptic plasticity to regional collaboration.

Biological neural networks are capable of flexibly modulating **synaptic plasticity** in response to dynamic inputs, including (1) stabilization of previously-learned synaptic changes to overcome subsequent interference [155], [482], [507], which motivates the strategies of weight regularization to selectively penalize parameter changes [12], [222], [497]; (2) number expansion and pruning of functional connections to provide flexibility for new memory formation [26], [100], [368], which derives the expansion-renormalization paradigm of creating extra space for learning the current task and renormalizing it with the previous one [382], [441], [442]; (3) activity-dependent and persistent regulation of synaptic plasticity, i.e., meta-plasticity or "plasticity of synaptic plasticity" [3], [4], [78], corresponding to the use of meta-learning [120], [186], [369]; and (4) inhibitory synapses for excited neurons to reduce the activity of other neurons [19], [54], [114], [179], which acts similarly to the binary mask for parameter allocation [384], [438].

As for **regional collaboration**, the *complementary learning system* (CLS) theory [233], [296] has been widely used to inspire continual learning, which attributes the advantages of biological learning and memory to the complementary functions of hippocampus and neocortex. The hippocampus is responsible for rapid acquisition of separated representations of specific experiences, while the neocortex enables progressive acquisition of structured knowledge for generalization. In particular, the hippocampus can replay neural representations of the previous experiences, so its function is often represented by a memory buffer [41], [373], [375], [444], [452] or a generative model [374], [392], [422], [438], [464] to recover the old data distributions. Instead of using the original data format, however, the hippocampal replay is in a temporally compressed form (e.g., an experience of around 6 seconds can be compressed into 0.2 seconds) [52], [85], [229], which potentially improves resource efficiency. Such biological advantage is further incorporated by replaying compressed data [444] or feature representations [25], [156]. In contrast, the neocortex-dependent memories are typically long-term and generalizable, where the representations can be progressively acquired in an unsupervised fashion [104], [105] and exhibit orthogonality [121], [474], consistent with the advantages of self-supervised and pre-trained representations in continual learning [168], [286], [300], [343], [360]. Besides, some regions of the biological brain have a modular architecture similar to the mixture of experts [184], such as the prefrontal cortex (i.e., a neocortex region) regulating sensory information from various cortical modules [124], [324], [421] and the mushroom body (i.e., a memory center in fruit fly) coordinating multiple parallel compartments to process the sequentially-arrived experiences [22], [23], [79], [314], [432], corresponding to the use of modular networks in continual learning [14], [361], [443].

8 Conclusion

In this work, we present an up-to-date and comprehensive survey of continual learning, bridging the latest advances in theory, method and application. We summarize both general objectives and particular challenges in this field, with an extensive analysis of how representative methods address them. Encouragingly, we observe a growing and widespread interest in continual learning from the broad AI community, bringing novel understandings, diversified applications and cross-directional opportunities. Based on such a holistic perspective, we expect the development of continual learning to eventually empower AI systems with human-like adaptability, responding flexibly to real-world dynamics and evolving themselves in a lifelong manner.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2020AAA0106000, 2020AAA0104304, 2020AAA0106302, 2021YFB2701000), NSFC Projects (62061136001, 62106123, 62076147, U19B2034, U1811461, U19A2081, 61972224), BNRist (BNR2022RC01006), the Tsinghua Institute for Guo Qiang, Tsinghua-OPPO Joint Research Center for Future Terminal Technology, and the High Performance Computing Center, Tsinghua University. L.W. was also supported by Shuimu Tsinghua Scholar.

References

[1] Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. Conditional channel gated networks for task-aware continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3931–3940, 2020.

[2] Mohamed Abdelsalam, Mojtaba Faramarzi, Shagun Sodhani, and Sarath Chandar. Iirc: Incremental implicitly-refined classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11038–11047, 2021.

[3] Wickliffe C Abraham. Metaplasticity: tuning synapses and networks for plasticity. *Nature Reviews Neuroscience*, 9(5):387–387, 2008.

[4] Wickliffe C Abraham and Mark F Bear. Metaplasticity: the plasticity of synaptic plasticity. *Trends in Neurosciences*, 19(4):126–130, 1996.

[5] Alessandro Achille, Tom Eccles, Loic Matthey, Chris Burgess, Nicholas Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. *Advances in Neural Information Processing Systems*, 31, 2018.

[6] Idan Achituve, Aviv Navon, Yochai Yemini, Gal Chechik, and Ethan Fetaya. Gp-tree: A gaussian process classifier for few-shot incremental learning. In *International Conference on Machine Learning*, pages 54–65. PMLR, 2021.

[7] Tameem Adel, Han Zhao, and Richard E Turner. Continual learning with adaptive weights (claw). In *International Conference on Learning Representations*, 2019.

[8] Aishwarya Agarwal, Biplab Banerjee, Fabio Cuzzolin, and Subhasis Chaudhuri. Semantics-driven generative replay for few-shot class incremental learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 5246–5254, 2022.

[9] Hongjoon Ahn, Sungmin Cha, Donggyu Lee, and Taesup Moon. Uncertainty-based continual learning with adaptive regularization. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 844–853, 2021.

[11] Afra Feyza Akyürek, Ekin Akyürek, Derry Wijaya, and Jacob Andreas. Subspace regularizers for few-shot class incremental learning. In *International Conference on Learning Representations*, 2021.

[12] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision*, pages 139–154, 2018.

[13] Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 11849–11860. Curran Associates, Inc., 2019.

[14] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3366–3375, 2017.

[15] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11263, 2019.

[16] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. In *International Conference on Learning Representations*, 2021.

[18] Randy Ardywibowo, Zepeng Huo, Zhangyang Wang, Bobak J Mortazavi, Shuai Huang, and Xiaoning Qian. Varigrow: Variational architecture growing for task-agnostic continual learning based on bayesian novelty. In *International Conference on Machine Learning*, pages 865–877. PMLR, 2022.

[19] Armen C Arevian, Vikrant Kapoor, and Nathaniel N Urban. Activity-dependent gating of lateral inhibition in the mouse olfactory bulb. *Nature Neuroscience*, 11(1):80–87, 2008.

[20] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.

[21] Arjun Ashok, KJ Joseph, and Vineeth N Balasubramanian. Class-incremental learning with cross-space clustering and controlled transfer. In *European Conference on Computer Vision*, pages 105–122. Springer, 2022.

[22] Yoshinori Aso and Gerald M Rubin. Dopaminergic neurons write and update memories with cell-type-specific rules. *Elife*, 5:e16135, 2016.

[23] Yoshinori Aso, Divya Sitaraman, Toshiharu Ichinose, Karla R Kaun, Katrin Vogt, Ghislain Belliart-Guérin, Pierre-Yves Plaçais, Alice A Robie, Nobuhiro Yamagata, Christopher Schnaitmann, et al. Mushroom body output neurons encode valence and guide memory-based action selection in drosophila. *Elife*, 3:e04580, 2014.

[24] Ali Ayub and Carter Fendley. Few-shot continual active learning by a robot. *arXiv preprint arXiv:2210.04137*, 2022.

[25] Ali Ayub and Alan Wagner. Eec: Learning to encode and regenerate images for continual learning. In *International Conference on Learning Representations*, 2020.

[26] Craig H Bailey, Eric R Kandel, and Kristen M Harris. Structural components of synaptic plasticity and memory consolidation. *Cold Spring Harbor Perspectives in Biology*, 7(7):a021758, 2015.

[27] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8218–8227, 2021.

[28] Jihwan Bang, Hyunseo Koh, Seulki Park, Hwanjun Song, Jung-Woo Ha, and Jonghyun Choi. 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*, pages 9275–9284, 2022.

[29] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2021.

[30] Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O Stanley, Jeff Clune, and Nick Cheney. Learning to continually learn. In *ECAI 2020*, pages 992–1001. IOS Press, 2020.

[31] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 583–592, 2019.

[32] Eden Belouadah and Adrian Popescu. Scail: Classifier weights scaling for class incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1266–1275, 2020.

[33] Mehdi Abbana Bennani, Thang Doan, and Masashi Sugiyama.

Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020.

[34] Frederik Benzing. Unifying importance based regularisation methods for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2372–2396. PMLR, 2022.

[35] Glen Berseth, Zhiwei Zhang, Grace Zhang, Chelsea Finn, and Sergey Levine. Comps: Continual meta policy search. In *International Conference on Learning Representations*, 2021.

[36] Prashant Shivaram Bhat, Bahram Zonooz, and Elahe Arani. Task-aware information routing from common representation space in lifelong learning. In *International Conference on Learning Representations*, 2023.

[37] Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Subhadeep Koley, Rohit Kundu, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Doodle it yourself: Class incremental learning by drawing a few sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2293–2302, 2022.

[38] Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-jussà. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, 2020.

[39] Lorenzo Bonicelli, Matteo Boschini, Angelo Porrello, Concetto Spampinato, and Simone Calderara. On the effectiveness of lipschitz-driven rehearsal in continual learning. *arXiv preprint arXiv:2210.06443*, 2022.

[40] Zalán Borsos, Mojmir Mutny, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. *Advances in Neural Information Processing Systems*, 33:14879–14890, 2020.

[41] Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended der-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[42] Matteo Boschini, Lorenzo Bonicelli, Angelo Porrello, Giovanni Bellitto, Matteo Pennisi, Simone Palazzo, Concetto Spampinato, and Simone Calderara. Transfer without forgetting. *arXiv preprint arXiv:2206.00388*, 2022.

[43] Matteo Boschini, Pietro Buzzega, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Continual semi-supervised learning through contrastive interpolation consistency. *Pattern Recognition Letters*, 162:9–14, 2022.

[44] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

[45] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[46] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in Neural Information Processing Systems*, 33:15920–15930, 2020.

[47] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations*, 2021.

[48] Lucas Caccia, Eugene Belilovsky, Massimo Caccia, and Joelle Pineau. Online learned continual compression with adaptive quantization modules. In *International Conference on Machine Learning*, pages 1240–1250. PMLR, 2020.

[49] Massimo Caccia, Pau Rodriguez, Oleksiy Ostapenko, Fabrice Normandin, Min Lin, Lucas Page-Caccia, Issam Hadj Laradji, Irina Rish, Alexandre Lacoste, David Vázquez, et al. Online fast adaptation and knowledge accumulation (osaka): a new approach to continual learning. *Advances in Neural Information Processing Systems*, 33:16532–16545, 2020.

[50] Zhipeng Cai, Ozan Sener, and Vladlen Koltun. Online continual learning with natural distribution shifts: An empirical study with visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8281–8290, 2021.

[51] Yue Cao, Hao-Ran Wei, Boxing Chen, and Xiaojun Wan. Continual learning for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3964–3974, 2021.

[52] Margaret F Carr, Shantanu P Jadhav, and Loren M Frank. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature neuroscience*, 14(2):147–153, 2011.

[53] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision*, pages 233–248, 2018.

[54] N Alex Cayco-Gajic and R Angus Silver. Re-evaluating circuit mechanisms underlying pattern separation. *Neuron*, 101(4):584–602, 2019.

[55] Fabio Cermelli, Dario Fontanel, Antonio Tavera, Marco Ciccone, and Barbara Caputo. Incremental learning in semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4371–4381, 2022.

[56] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9233–9242, 2020.

[57] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9516–9525, 2021.

[58] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

[59] Sungmin Cha, Soonwon Hong, Moontae Lee, and Taesup Moon. Task-balanced batch normalization for exemplar-based class-incremental learning. *arXiv preprint arXiv:2201.12559*, 2022.

[60] Sungmin Cha, Hsiang Hsu, Taebaek Hwang, Flavio Calmon, and Taesup Moon. Cpr: Classifier-projection regularization for continual learning. In *International Conference on Learning Representations*, 2020.

[61] Sungmin Cha, YoungJoon Yoo, Taesup Moon, et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *Advances in Neural Information Processing Systems*, 34:10919–10930, 2021.

[62] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30, 2017.

[63] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision*, pages 532–547, 2018.

[64] Arslan Chaudhry, Albert Gordo, Puneet Dokania, Philip Torr, and David Lopez-Paz. Using hindsight to anchor past knowledge in continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6993–7001, 2021.

[65] Arslan Chaudhry, Naeemullah Khan, Puneet Dokania, and Philip Torr. Continual learning in low-rank orthogonal subspaces. *Advances in Neural Information Processing Systems*, 33:9900–9911, 2020.

[66] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2018.

[67] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

[68] Cheng Chen, Ji Zhang, Jingkuan Song, and Lianli Gao. Class gradient projection for continual learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 5575–5583, 2022.

[69] Hung-Jen Chen, An-Chieh Cheng, Da-Cheng Juan, Wei Wei, and Min Sun. Mitigating forgetting in online continual learning via instance-aware parameterization. *Advances in Neural Information Processing Systems*, 33:17466–17477, 2020.

[70] Pei Chen, Yangkang Zhang, Zejian Li, and Lingyun Sun. Few-shot incremental learning for label-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3697–3707, 2022.

[71] Zhiyuan Chen and Bing Liu. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207, 2018.

[72] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2534–2543, 2021.

[73] Ali Cheraghian, Shafin Rahman, Sameera Ramasinghe, Pengfei

Fang, Christian Simon, Lars Petersson, and Mehrtash Harandi. Synthesized feature based few-shot class-incremental learning on a mixture of subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8661–8670, 2021.
[74] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14166–14175, 2022.
[75] Ron Chrisley. Embodied artificial intelligence. *Artificial intelligence*, 149(1):131–150, 2003.
[76] Aristotelis Chrysakis and Marie-Francine Moens. Online continual learning from imbalanced data. In *International Conference on Machine Learning*, pages 1952–1961. PMLR, 2020.
[77] Nikhil Churamani, Sinan Kalkan, and Hatice Gunes. Continual learning for affective robotics: Why, what and how? In *IEEE International Conference on Robot and Human Interactive Communication*, pages 425–431. IEEE, 2020.
[78] Roger L Clem, Tansu Celikel, and Alison L Barth. Ongoing in vivo experience triggers synaptic metaplasticity in the neocortex. *Science*, 319(5859):101–104, 2008.
[79] Raphael Cohn, Ianessa Morantte, and Vanessa Ruta. Coordinated and compartmentalized neuromodulation shapes sensory processing in drosophila. *Cell*, 163(7):1742–1755, 2015.
[80] Mark Collier, Efi Kokiopoulou, Andrea Gesmundo, and Jesse Berent. Routing networks with co-training for continual learning. *arXiv preprint arXiv:2009.04381*, 2020.
[81] Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. Gan memory with no forgetting. *Advances in Neural Information Processing Systems*, 33:16481–16494, 2020.
[82] Andrea Cossu, Tinne Tuytelaars, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, and Davide Bacciu. Continual pre-training mitigates forgetting in language and vision. *arXiv preprint arXiv:2205.09357*, 2022.
[83] Francesco D'Angelo and Christian Henning. Uncertainty-based out-of-distribution detection requires suitable function space priors. *arXiv preprint arXiv:2110.06020*, 2021.
[84] MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16712–16721, 2022.
[85] Thomas J Davidson, Fabian Kloosterman, and Matthew A Wilson. Hippocampal replay of extended experience. *Neuron*, 63(4):497–507, 2009.
[86] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2021.
[87] Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8250–8259, 2021.
[88] Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32, 2019.
[89] Riccardo Del Chiaro, Bartłomiej Twardowski, Andrew Bagdanov, and Joost Van de Weijer. Ratt: Recurrent attention to transient tasks for continual image captioning. *Advances in Neural Information Processing Systems*, 33:16736–16748, 2020.
[90] Danruo Deng, Guangyong Chen, Jianye Hao, Qiong Wang, and Pheng-Ann Heng. Flattening sharpness for dynamic gradient projection memory benefits continual learning. *Advances in Neural Information Processing Systems*, 34:18710–18721, 2021.
[91] Mohammad Mahdi Derakhshani, Xiantong Zhen, Ling Shao, and Cees Snoek. Kernel continual learning. In *International Conference on Machine Learning*, pages 2621–2631. PMLR, 2021.
[92] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
[93] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5138–5146, 2019.
[94] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
[95] Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR, 2021.
[96] Thang Doan, Seyed Iman Mirzadeh, Joelle Pineau, and Mehrdad Farajtabar. Efficient continual learning ensembles in neural network subspaces. *arXiv preprint arXiv:2202.09826*, 2022.
[97] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10164–10173, 2022.
[98] Na Dong, Yongqiang Zhang, Mingli Ding, and Gim Hee Lee. Bridging non co-occurrence with unlabeled in-the-wild data for incremental object detection. *Advances in Neural Information Processing Systems*, 34:30492–30503, 2021.
[99] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1255–1263, 2021.
[100] Tao Dong, Jing He, Shiqing Wang, Lianzhang Wang, Yuqi Cheng, and Yi Zhong. Inability to activate rac1-dependent forgetting contributes to behavioral inflexibility in mutants of multiple autism-risk genes. *Proceedings of the National Academy of Sciences*, 113(27):7644–7649, 2016.
[101] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4040–4050, 2021.
[102] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision*, pages 86–102. Springer, 2020.
[103] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9285–9295, 2022.
[104] Kenji Doya. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, 12(7-8):961–974, 1999.
[105] Kenji Doya. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current opinion in neurobiology*, 10(6):732–739, 2000.
[106] Fei Du, Yun Yang, Ziyuan Zhao, and Zeng Zeng. Efficient perturbation inference and expandable network for continual learning. *Neural Networks*, 2022.
[107] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022.
[108] Sayna Ebrahimi, Mohamed Elhoseiny, Trevor Darrell, and Marcus Rohrbach. Uncertainty-guided continual learning with bayesian neural networks. In *International Conference on Learning Representations*, 2019.
[109] Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. Adversarial continual learning. In *European Conference on Computer Vision*, pages 386–402. Springer, 2020.
[110] Sayna Ebrahimi, Suzanne Petryk, Akash Gokul, William Gan, Joseph E Gonzalez, Marcus Rohrbach, et al. Remembering for the right reasons: Explanations reduce catastrophic forgetting. In *International Conference on Learning Representations*, 2020.
[111] Avia Efrat and Omer Levy. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*, 2020.
[112] Evgenii Egorov, Anna Kuzina, and Evgeny Burnaev. Boovae: Boosting approach for continual learning of vae. *Advances in Neural Information Processing Systems*, 34:17889–17901, 2021.
[113] Beyza Ermis, Giovanni Zappella, Martin Wistuba, Aditya Rawal, and Cedric Archambeau. Memory efficient continual learning with transformers. *Advances in Neural Information Processing Systems*, 35:10629–10642, 2022.
[114] Claudia Espinoza, Segundo Jose Guzman, Xiaomin Zhang, and Peter Jonas. Parvalbumin+ interneurons obey unique connectivity rules and establish a powerful lateral-inhibition microcircuit in dentate gyrus. *Nature Communications*, 9(1):1–10, 2018.
[115] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3762–3773.

PMLR, 2020.
[116] Tao Feng, Mang Wang, and Hangjie Yuan. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9427–9436, 2022.
[117] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.
[118] Enrico Fini, Victor G Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2022.
[119] Enrico Fini, Stéphane Lathuiliere, Enver Sangineto, Moin Nabi, and Elisa Ricci. Online continual learning under extreme memory constraints. In *European Conference on Computer Vision*, pages 720–735. Springer, 2020.
[120] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.
[121] Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7):1258–1270, 2022.
[122] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.
[123] Mathieu Pagé Fortin and Brahim Chaib-draa. Continual semantic segmentation leveraging image-level labels and rehearsal. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 2, 2022.
[124] Paul W Frankland and Bruno Bontempi. The organization of recent and remote memories. *Nature Reviews Neuroscience*, 6(2):119–130, 2005.
[125] Stan Franklin. Autonomous agents as embodied ai. *Cybernetics & Systems*, 28(6):499–520, 1997.
[126] Samir Yitzhak Gadre, Kiana Ehsani, Shuran Song, and Roozbeh Mottaghi. Continuous scene representations for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14849–14859, 2022.
[127] Jhair Gallardo, Tyler L Hayes, and Christopher Kanan. Self-supervised training enhances online continual learning. *arXiv preprint arXiv:2103.14010*, 2021.
[128] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. Incremental few-shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1185–1194, 2021.
[129] Qiankun Gao, Chen Zhao, Bernard Ghanem, and Jian Zhang. R-dfcil: Relation-guided representation learning for data-free class incremental learning. *arXiv preprint arXiv:2203.13104*, 2022.
[130] Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, 2021.
[131] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in Neural Information Processing Systems*, 31, 2018.
[132] Christian Geishauser, Carel van Niekerk, Hsien-Chin Lin, Nurul Lubis, Michael Heck, Shutong Feng, and Milica Gasic. Dynamic dialogue policy for continual reinforcement learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 266–284, 2022.
[133] Binzong Geng, Fajie Yuan, Qiancheng Xu, Ying Shen, Ruifeng Xu, and Min Yang. Continual learning for task-oriented dialogue system with iterative network pruning, expanding and masking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 517–523, 2021.
[134] Subhankar Ghosh. Dynamic vaes with generative replay for continual zero-shot learning. *arXiv preprint arXiv:2104.12468*, 2021.
[135] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.
[136] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 792–801, 2021.
[137] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020.
[138] Siavash Golkar, Michael Kagan, and Kyunghyun Cho. Continual learning via neural pruning. *arXiv preprint arXiv:1903.04476*, 2019.
[139] Zheng Gong, Kun Zhou, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. Continual pre-training of language models for math problem understanding with syntax-aware memory network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5923–5933, 2022.
[140] Saisubramaniam Gopalakrishnan, Pranshu Ranjan Singh, Haytham Fayek, Savitha Ramasamy, and Arulmurugan Ambikapathi. Knowledge capture and replay for continual learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 10–18, 2022.
[141] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.
[142] Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605, 2019.
[143] Shuhao Gu and Yang Feng. Investigating catastrophic forgetting during continual training for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4315–4326, 2020.
[144] Yanan Gu, Cheng Deng, and Kun Wei. Class-incremental instance segmentation via multi-teacher networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1478–1486, 2021.
[145] Yanan Gu, Xu Yang, Kun Wei, and Cheng Deng. Not just selection, but exploration: Online class-incremental continual learning via dual view consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7442–7451, 2022.
[146] Yiduo Guo, Wenpeng Hu, Dongyan Zhao, and Bing Liu. Adaptive orthogonal projection for batch and online continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 2, 2022.
[147] Yiduo Guo, Bing Liu, and Dongyan Zhao. Online continual learning through mutual information maximization. In *International Conference on Machine Learning*, pages 8109–8126. PMLR, 2022.
[148] Gunshi Gupta, Karmesh Yadav, and Liam Paull. Look-ahead meta learning for continual learning. *Advances in Neural Information Processing Systems*, 33:11588–11598, 2020.
[149] Mustafa B Gurbuz and Constantine Dovrolis. Nispa: Neuro-inspired stability-plasticity adaptation for continual learning in sparse networks. In *International Conference on Machine Learning*, pages 8157–8174. PMLR, 2022.
[150] Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020.
[151] Rujun Han, Xiang Ren, and Nanyun Peng. Econet: Effective continual pretraining of language models for event temporal reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5367–5380, 2021.
[152] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4143–4152, 2019.
[153] Yu Hao, Yanwei Fu, Yu-Gang Jiang, and Qi Tian. An end-to-end architecture for class-incremental object detection with knowledge distillation. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2019.
[154] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
[155] Akiko Hayashi-Takagi, Sho Yagishita, Mayumi Nakamura, Fukutoshi Shirai, Yi I Wu, Amanda L Loshbaugh, Brian Kuhlman, Klaus M Hahn, and Haruo Kasai. Labelling and optical erasure of synaptic memory traces in the motor cortex. *Nature*, 525(7569):333–338, 2015.
[156] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent

catastrophic forgetting. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020.
[157] Tyler L Hayes, Giri P Krishnan, Maxim Bazhenov, Hava T Siegelmann, Terrence J Sejnowski, and Christopher Kanan. Replay in deep learning: Current approaches and missing biological elements. *Neural Computation*, 33(11):2908–2950, 2021.
[158] Chen He, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exemplar-supported generative reproduction for class incremental learning. In *BMVC*, page 98, 2018.
[159] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu. Incremental learning in online scenario. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13926–13935, 2020.
[160] Jiangpeng He and Fengqing Zhu. Online continual learning via candidates voting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3154–3163, 2022.
[161] Christian Henning, Maria Cervera, Francesco D'Angelo, Johannes Von Oswald, Regina Traber, Benjamin Ehret, Seijin Kobayashi, Benjamin F Grewe, and João Sacramento. Posterior meta-replay for continual learning. *Advances in Neural Information Processing Systems*, 34:14135–14149, 2021.
[162] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9057–9067, 2022.
[163] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.
[164] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Lifelong learning via progressive distillation and retrospection. In *Proceedings of the European Conference on Computer Vision*, pages 437–452, 2018.
[165] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019.
[166] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.
[167] Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018.
[168] Dapeng Hu, Shipeng Yan, Qizhengqiu Lu, HONG Lanqing, Hailin Hu, Yifan Zhang, Zhenguo Li, Xinchao Wang, and Jiashi Feng. How well does self-supervised pre-training perform with streaming data? In *International Conference on Learning Representations*, 2021.
[169] Huiyi Hu, Ang Li, Daniele Calandriello, and Dilan Gorur. One pass imagenet. *arXiv preprint arXiv:2111.01956*, 2021.
[170] Wenpeng Hu, Zhou Lin, Bing Liu, Chongyang Tao, Zhengwei Tao Tao, Dongyan Zhao, Jinwen Ma, and Rui Yan. Overcoming catastrophic forgetting for continual learning via model adaptation. In *International Conference on Learning Representations*, 2019.
[171] Wenpeng Hu, Qi Qin, Mengyu Wang, Jinwen Ma, and Bing Liu. Continual learning by using information of each class holistically. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7797–7805, 2021.
[172] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3957–3966, 2021.
[173] Yufan Huang, Yanzhe Zhang, Jiaao Chen, Xuezhi Wang, and Diyi Yang. Continual learning for text classification with information disentanglement based regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2736–2746, 2021.
[174] Zhongzhan Huang, Mingfu Liang, Senwei Liang, and Wei He. Altersgd: Finding flat minima for continual learning by alternative training. *arXiv preprint arXiv:2107.05804*, 2021.
[175] Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, picking and growing for unforgetting continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.
[176] Julio Hurtado, Alain Raymond, and Alvaro Soto. Optimizing reusable knowledge for continual learning via metalearning. *Advances in Neural Information Processing Systems*, 34:14150–14162, 2021.
[177] Ferenc Huszár. On quadratic penalties in elastic weight consolidation. *arXiv preprint arXiv:1712.03847*, 2017.
[178] Rakib Hyder, Ken Shao, Boyu Hou, Panos Markopoulos, Ashley Prater-Bennette, and M Salman Asif. Incremental task learning with incremental rank updates. In *European Conference on Computer Vision*, pages 566–582. Springer, 2022.
[179] Kengo Inada, Yoshiko Tsuchimoto, and Hokto Kazama. Origins of cell-type-specific olfactory processing in the drosophila mushroom body circuit. *Neuron*, 95(2):357–367, 2017.
[180] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015.
[181] Ahmet Iscen, Jeffrey Zhang, Svetlana Lazebnik, and Cordelia Schmid. Memory-efficient incremental learning through feature adaptation. In *European Conference on Computer Vision*, pages 699–715. Springer, 2020.
[182] David Isele and Akansel Cosgun. Selective experience replay for lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
[183] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.
[184] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.
[185] Paul Janson, Wenxuan Zhang, Rahaf Aljundi, and Mohamed Elhoseiny. A simple baseline that questions the use of pretrained-models in continual learning. In *NeurIPS Workshops*, 2022.
[186] Khurram Javed and Martha White. Meta-learning representations for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.
[187] Peter Jedlicka, Matus Tomko, Anthony Robins, and Wickliffe C Abraham. Contributions by metaplasticity to solving the catastrophic forgetting problem. *Trends in Neurosciences*, 2022.
[188] Ghassen Jerfel, Erin Grant, Tom Griffiths, and Katherine A Heller. Reconciling meta-learning and continual learning with online mixtures of tasks. *Advances in Neural Information Processing Systems*, 32, 2019.
[189] Ziyue Jiang, Yi Ren, Ming Lei, and Zhou Zhao. Fedspeech: Federated text-to-speech with continual learning. *arXiv preprint arXiv:2110.07216*, 2021.
[190] Hyundong Jin and Eunwoo Kim. Helpful or harmful: Inter-task association in continual learning. In *European Conference on Computer Vision*, pages 519–535. Springer, 2022.
[191] Xisen Jin, Bill Yuchen Lin, Mohammad Rostami, and Xiang Ren. Learn continually, generalize rapidly: Lifelong knowledge accumulation for few-shot learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 714–729, 2021.
[192] Xisen Jin, Arka Sadhu, Junyi Du, and Xiang Ren. Gradient-based editing of memory examples for online task-free continual learning. *Advances in Neural Information Processing Systems*, 34:29193–29205, 2021.
[193] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Vineeth N Balasubramanian. Energy-based latent aligner for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7452–7461, 2022.
[194] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021.
[195] KJ Joseph, Sujoy Paul, Gaurav Aggarwal, Soma Biswas, Piyush Rai, Kai Han, and Vineeth N Balasubramanian. Novel class discovery without forgetting. In *European Conference on Computer Vision*, pages 570–586. Springer, 2022.
[196] Sangwon Jung, Hongjoon Ahn, Sungmin Cha, and Taesup Moon. Continual learning with node-importance based adaptive group sparse regularization. *Advances in Neural Information Processing Systems*, 33:3647–3658, 2020.
[197] Jayateja Kalla and Soma Biswas. S3c: Self-supervised stochastic classifiers for few-shot class-incremental learning. In *European Conference on Computer Vision*, pages 432–448. Springer, 2022.
[198] Menelaos Kanakis, David Bruggemann, Suman Saha, Stamatios Georgoulis, Anton Obukhov, and Luc Van Gool. Reparameterizing convolutions for incremental multi-task learning without task interference. In *European Conference on Computer Vision*, pages

689–707. Springer, 2020.
[199] Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, and Chang D Yoo. Forget-free continual learning with winning subnetworks. In *International Conference on Machine Learning*, pages 10734–10750. PMLR, 2022.
[200] Minsoo Kang, Jaeyoo Park, and Bohyung Han. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16071–16080, 2022.
[201] Zhiqi Kang, Enrico Fini, Moin Nabi, Elisa Ricci, and Karteek Alahari. A soft nearest-neighbor framework for continual semi-supervised learning. *arXiv preprint arXiv:2212.05102*, 2022.
[202] Ta-Chu Kao, Kristopher Jensen, Gido van de Ven, Alberto Bernacchia, and Guillaume Hennequin. Natural continual learning: success is a journey, not (just) a destination. *Advances in Neural Information Processing Systems*, 34:28067–28079, 2021.
[203] Christos Kaplanis, Claudia Clopath, and Murray Shanahan. Continual reinforcement learning with multi-timescale replay. *arXiv preprint arXiv:2004.07530*, 2020.
[204] Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Continual reinforcement learning with complex synapses. In *International Conference on Machine Learning*, pages 2497–2506. PMLR, 2018.
[205] Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Policy consolidation for continual reinforcement learning. In *International Conference on Machine Learning*, pages 3242–3251. PMLR, 2019.
[206] Sanyam Kapoor, Theofanis Karaletsos, and Thang D Bui. Variational auto-regressive gaussian processes for continual learning. In *International Conference on Machine Learning*, pages 5290–5300. PMLR, 2021.
[207] Ryo Karakida and Shotaro Akaho. Learning curves for continual learning in neural networks: Self-knowledge transfer and forgetting. In *International Conference on Learning Representations*, 2022.
[208] Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. Continual training of language models for few-shot learning. *arXiv preprint arXiv:2210.05549*, 2022.
[209] Zixuan Ke and Bing Liu. Continual learning of natural language processing tasks: A survey. *arXiv preprint arXiv:2211.12701*, 2022.
[210] Zixuan Ke, Bing Liu, and Xingchang Huang. Continual learning of a mixed sequence of similar and dissimilar tasks. *Advances in Neural Information Processing Systems*, 33:18493–18504, 2020.
[211] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. In *International Conference on Learning Representations*, 2018.
[212] Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.
[213] Samuel Kessler, Jack Parker-Holder, Philip Ball, Stefan Zohren, and Stephen J Roberts. Same state, different task: Continual reinforcement learning without interference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7143–7151, 2022.
[214] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.
[215] Mert Kilickaya, Joost Van der Weijer, and Yuki Asano. Towards label-efficient incremental learning: A survey. *arXiv preprint arXiv:2302.00353*, 2023.
[216] Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim. Imbalanced continual learning with partitioning reservoir sampling. In *European Conference on Computer Vision*, pages 411–428. Springer, 2020.
[217] Chris Dongjoo Kim, Jinseo Jeong, Sangwoo Moon, and Gunhee Kim. Continual learning on noisy data streams via self-purified replay. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 537–547, 2021.
[218] Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. *arXiv preprint arXiv:2211.02633*, 2022.
[219] Jong-Yeong Kim and Dong-Wan Choi. Split-and-bridge: Adaptable class incremental learning within a single neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8137–8145, 2021.
[220] Seoyoon Kim, Seongjun Yun, and Jaewoo Kang. Dygrain: An incremental learning framework for dynamic graphs. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3157–3163, 2022.
[221] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
[222] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.
[223] Joseph KJ and Vineeth N Balasubramanian. Meta-consolidation for continual learning. *Advances in Neural Information Processing Systems*, 33:14374–14386, 2020.
[224] Joseph Kj, Jathushan Rajasegaran, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Incremental object detection via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
[225] Jeremias Knoblauch, Hisham Husain, and Tom Diethe. Optimal continual learning has perfect memory and is np-hard. In *International Conference on Machine Learning*, pages 5327–5337. PMLR, 2020.
[226] Hyunseo Koh, Dahyun Kim, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on class incremental blurry task configuration with anytime inference. In *International Conference on Learning Representations*, 2021.
[227] Yajing Kong, Liu Liu, Zhen Wang, and Dacheng Tao. Balancing stability and plasticity through advanced null space in continual learning. In *European Conference on Computer Vision*, pages 219–236. Springer, 2022.
[228] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.
[229] Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, et al. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210, 2022.
[230] Anna Kukleva, Hilde Kuehne, and Bernt Schiele. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9020–9029, 2021.
[231] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.
[232] Abhishek Kumar, Sunabha Chatterjee, and Piyush Rai. Bayesian structural adaptation for continual learning. In *International Conference on Machine Learning*, pages 5850–5860. PMLR, 2021.
[233] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.
[234] Lilly Kumari, Shengjie Wang, Tianyi Zhou, and Jeff A Bilmes. Retrospective adversarial replay for continual learning. *Advances in Neural Information Processing Systems*, 35:28530–28544, 2022.
[235] Richard Kurle, Botond Cseke, Alexej Klushyn, Patrick Van Der Smagt, and Stephan Günnemann. Continual learning with bayesian neural networks for non-stationary data. In *International Conference on Learning Representations*, 2019.
[236] Vinod K Kurmi, Badri N Patro, Venkatesh K Subramanian, and Vinay P Namboodiri. Do not forget to attend to uncertainty while mitigating catastrophic forgetting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 736–745, 2021.
[237] Alexis Lechat, Stéphane Herbin, and Frédéric Jurie. Semi-supervised class incremental learning. In *International Conference on Pattern Recognition*, pages 10383–10389. IEEE, 2021.
[238] Eugene Lee, Cheng-Han Huang, and Chen-Yi Lee. Few-shot and continual learning with attentive independent mechanisms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9455–9464, 2021.
[239] Janghyeon Lee, Hyeong Gwon Hong, Donggyu Joo, and Junmo Kim. Continual learning with extended kronecker-factored approximate curvature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9001–9010, 2020.
[240] Janghyeon Lee, Donggyu Joo, Hyeong Gwon Hong, and Junmo Kim. Residual continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4553–4560, 2020.
[241] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 312–321, 2019.

[242] Kwot Sin Lee, Ngoc-Trung Tran, and Ngai-Man Cheung. Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3942–3952, 2021.

[243] Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pages 6109–6119. PMLR, 2021.

[244] Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim. A neural dirichlet process mixture model for task-free continual learning. In *International Conference on Learning Representations*, 2019.

[245] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. *Advances in Neural Information Processing Systems*, 30, 2017.

[246] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information Fusion*, 58:52–68, 2020.

[247] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

[248] Aodong Li, Alex Boyd, Padhraic Smyth, and Stephan Mandt. Variational beam search for learning with distribution shifts. *arXiv preprint arXiv:2012.08101*, 2020.

[249] Dingcheng Li, Zheng Chen, Eunah Cho, Jie Hao, Xiaohu Liu, Fan Xing, Chenlei Guo, and Yang Liu. Overcoming catastrophic forgetting during domain adaptation of seq2seq language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5441–5454, 2022.

[250] Dawei Li, Serafettin Tasci, Shalini Ghosh, Jingwen Zhu, Junting Zhang, and Larry Heck. Rilod: Near real-time incremental learning for object detection at the edge. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pages 113–126, 2019.

[251] Guodun Li, Yuchen Zhai, Qianglong Chen, Xing Gao, Ji Zhang, and Yin Zhang. Continual few-shot intent detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 333–343, 2022.

[252] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[253] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020.

[254] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International Conference on Machine Learning*, pages 3925–3934. PMLR, 2019.

[255] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

[256] Zhuang Li, Lizhen Qu, and Gholamreza Haffari. Total recall: a customized continual learning method for neural semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3816–3831, 2021.

[257] Mingfu Liang, Jiahuan Zhou, Wei Wei, and Ying Wu. Balancing between forgetting and acquisition in incremental subpopulation learning. In *European Conference on Computer Vision*, pages 364–380. Springer, 2022.

[258] Guoliang Lin, Hanlu Chu, and Hanjiang Lai. Towards better plasticity-stability trade-off in incremental learning: A simple linear connector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2022.

[259] Hongbin Lin, Yifan Zhang, Zhen Qiu, Shuaicheng Niu, Chuang Gan, Yanxia Liu, and Mingkui Tan. Prototype-guided continual adaptation for class-incremental unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 351–368. Springer, 2022.

[260] Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Trgp: Trust region gradient projection for continual learning. In *International Conference on Learning Representations*, 2021.

[261] Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Beyond not-forgetting: Continual learning with backward knowledge transfer. *arXiv preprint arXiv:2211.00789*, 2022.

[262] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[263] Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, et al. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *arXiv preprint arXiv:2309.06256*, 2023.

[264] Bing Liu and Sahisnu Mazumder. Lifelong and continual learning dialogue systems: learning during conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15058–15063, 2021.

[265] Huan Liu, Li Gu, Zhixiang Chi, Yang Wang, Yuanhao Yu, Jun Chen, and Jin Tang. Few-shot class-incremental learning via entropy-regularized data-free replay. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

[266] Hao Liu and Huaping Liu. Continual learning with recursive gradient optimization. In *International Conference on Learning Representations*, 2021.

[267] Minqian Liu, Shiyu Chang, and Lifu Huang. Incremental prompting: Episodic memory prompt for lifelong event detection. *arXiv preprint arXiv:2204.07275*, 2022.

[268] Qing Liu, Orchid Majumder, Alessandro Achille, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Incremental meta-learning via indirect discriminant alignment. *arXiv preprint arXiv:2002.04162*, 2020.

[269] Qingbin Liu, Xiaoyan Yu, Shizhu He, Kang Liu, and Jun Zhao. Lifelong intent detection via multi-strategy rebalancing. *arXiv preprint arXiv:2108.04445*, 2021.

[270] Tianlin Liu, Lyle Ungar, and Joao Sedoc. Continual learning for sentence representations using conceptors. In *Proceedings of NAACL-HLT*, pages 3274–3279, 2019.

[271] Xialei Liu, Yu-Song Hu, Xu-Sheng Cao, Andrew D Bagdanov, Ke Li, and Ming-Ming Cheng. Long-tailed class incremental learning. In *European Conference on Computer Vision*, pages 495–512. Springer, 2022.

[272] Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *International Conference on Pattern Recognition*, pages 2262–2268. IEEE, 2018.

[273] Xialei Liu, Chenshen Wu, Mikel Menta, Luis Herranz, Bogdan Raducanu, Andrew D Bagdanov, Shangling Jui, and Joost van de Weijer. Generative feature replay for class-incremental learning. In *CVPR Workshops*, pages 226–227, 2020.

[274] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2544–2553, 2021.

[275] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Rmm: Reinforced memory management for class-incremental learning. *Advances in Neural Information Processing Systems*, 34:3478–3490, 2021.

[276] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12245–12254, 2020.

[277] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[278] Vincenzo Lomonaco, Karan Desai, Eugenio Culurciello, and Davide Maltoni. Continual reinforcement learning in 3d non-stationary environments. In *CVPR Workshops*, pages 248–249, 2020.

[279] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26. PMLR, 2017.

[280] Noel Loo, Siddharth Swaroop, and Richard E Turner. Generalized variational continual learning. In *International Conference on Learning Representations*, 2020.

[281] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[282] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion

probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, 2022.

[283] Kevin Lu, Igor Mordatch, and Pieter Abbeel. Adaptive online planning for continual lifelong learning. *arXiv preprint arXiv:1912.01188*, 2019.

[284] Yilin Lyu, Liyuan Wang, Xingxing Zhang, Zicheng Sun, Hang Su, Jun Zhu, and Liping Jing. Overcoming recency bias of normalization statistics in continual learning: Balance and adaptation. *arXiv preprint arXiv:2310.08855*, 2023.

[285] Yuhang Ma, Zhongle Xie, Jue Wang, Ke Chen, and Lidan Shou. Continual federated learning based on knowledge distillation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 3, 2022.

[286] Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations*, 2021.

[287] Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul A Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, 2021.

[288] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.

[289] Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3589–3599, 2021.

[290] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision*, pages 67–82, 2018.

[291] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

[292] Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. Recall: Replay-based continual learning in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7026–7035, 2021.

[293] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pages 2408–2417. PMLR, 2015.

[294] Marc Masana, Bartłomiej Twardowski, and Joost Van de Weijer. On class orderings for incremental learning. *arXiv preprint arXiv:2007.02145*, 2020.

[295] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2337–2345, 2021.

[296] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

[297] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[298] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[299] Nikhil Mehta, Kevin Liang, Vinay Kumar Verma, and Lawrence Carin. Continual learning using a bayesian nonparametric dictionary of weight factors. In *International Conference on Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2021.

[300] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *arXiv preprint arXiv:2112.09153*, 2021.

[301] Jorge Mendez, Boyu Wang, and Eric Eaton. Lifelong policy gradient learning of factored policies for faster training without forgetting. *Advances in Neural Information Processing Systems*, 33:14398–14409, 2020.

[302] Fei Mi, Liangwei Chen, Mengjie Zhao, Minlie Huang, and Boi Faltings. Continual learning for natural language generation in task-oriented dialog systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3461–3474, 2020.

[303] Fei Mi, Lingjing Kong, Tao Lin, Kaicheng Yu, and Boi Faltings. Generalized class incremental learning. In *CVPR Workshops*, pages 240–241, 2020.

[304] Zichen Miao, Ze Wang, Wei Chen, and Qiang Qiu. Continual learning with filter atom swapping. In *International Conference on Learning Representations*, 2021.

[305] Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. In *ICCV Workshops*, pages 0–0, 2019.

[306] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1114–1124, 2021.

[307] Umberto Michieli and Pietro Zanuttigh. Knowledge distillation for incremental learning in semantic segmentation. *Computer Vision and Image Understanding*, 205:103167, 2021.

[308] Youngjae Min, Kwangjun Ahn, and Navid Azizan. One-pass learning via bridging orthogonal gradient descent and recursive least-squares. *arXiv preprint arXiv:2207.13853*, 2022.

[309] Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. In *International Conference on Machine Learning*, pages 15699–15717. PMLR, 2022.

[310] Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Timothy Nguyen, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Architecture matters in continual learning. *arXiv preprint arXiv:2202.00275*, 2022.

[311] Seyed Iman Mirzadeh, Mehrdad Farajtabar, and Hassan Ghasemzadeh. Dropout as an implicit gating mechanism for continual learning. In *CVPR Workshops*, pages 232–233, 2020.

[312] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *International Conference on Learning Representations*, 2020.

[313] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. *Advances in Neural Information Processing Systems*, 33:7308–7320, 2020.

[314] Mehrab N Modi, Yichun Shuai, and Glenn C Turner. The drosophila mushroom body: From architecture to algorithm in a learning circuit. *Annual Review of Neuroscience*, 43:465–484, 2020.

[315] Natawut Monaikul, Giuseppe Castellucci, Simone Filice, and Oleg Rokhlenko. Continual learning for named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13570–13577, 2021.

[316] Hadi Nekoei, Akilesh Badrinaaraayanan, Aaron Courville, and Sarath Chandar. Continuous coordination as a realistic scenario for lifelong learning. In *International Conference on Machine Learning*, pages 8016–8024. PMLR, 2021.

[317] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in Neural Information Processing Systems*, 33:512–523, 2020.

[318] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

[319] Youngmin Oh, Donghyeon Baek, and Bumsub Ham. Alife: Adaptive logit regularizer and feature replay for incremental semantic segmentation. *arXiv preprint arXiv:2210.06816*, 2022.

[320] OpenAI. Gpt-4 technical report. 2023.

[321] Oleksiy Ostapenko, Timothee Lesort, Pau Rodríguez, Md Rifat Arefin, Arthur Douillard, Irina Rish, and Laurent Charlin. Foundational models for continual learning: An empirical study of latent replay. *arXiv preprint arXiv:2205.00329*, 2022.

[322] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11321–11329, 2019.

[323] Oleksiy Ostapenko, Pau Rodriguez, Massimo Caccia, and Laurent Charlin. Continual learning via local module composition. *Advances in Neural Information Processing Systems*, 34:30298–30312, 2021.

[324] Torben Ott and Andreas Nieder. Dopamine and cognitive control in prefrontal cortex. *Trends in Cognitive Sciences*, 23(3):213–234, 2019.

[325] Inyoung Paik, Sangjun Oh, Taeyeong Kwak, and Injung Kim. Overcoming catastrophic forgetting by neuron-level plasticity control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5339–5346, 2020.

[326] Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschen-

hagen, Richard Turner, and Mohammad Emtiyaz E Khan. Continual deep learning by functional regularisation of memorable past. *Advances in Neural Information Processing Systems*, 33:4453–4464, 2020.

[327] Aristeidis Panos, Yuriko Kobe, Daniel Olmeda Reino, Rahaf Aljundi, and Richard E Turner. First session adaptation: A strong replay-free baseline for class-incremental learning. *arXiv preprint arXiv:2303.13199*, 2023.

[328] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[329] Dongmin Park, Seokil Hong, Bohyung Han, and Kyoung Mu Lee. Continual learning by asymmetric loss approximation with single-side overestimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3335–3344, 2019.

[330] Jaeyoo Park, Minsoo Kang, and Bohyung Han. Class-incremental learning for action recognition in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13698–13707, 2021.

[331] Ramakanth Pasunuru, Veselin Stoyanov, and Mohit Bansal. Continual few-shot learning for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5688–5702, 2021.

[332] Francesco Pelosin. Simpler is better: off-the-shelf continual learning through pretrained backbones. *arXiv preprint arXiv:2205.01586*, 2022.

[333] Binghui Peng and Andrej Risteski. Continual learning: a feature extraction formalization, an efficient algorithm, and fundamental obstructions. *arXiv preprint arXiv:2203.14383*, 2022.

[334] Can Peng, Kun Zhao, and Brian C Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. *Pattern Recognition Letters*, 140:109–115, 2020.

[335] Can Peng, Kun Zhao, Sam Maksoud, Meng Li, and Brian C Lovell. Sid: Incremental learning for anchor-free object detection via selective and inter-related distillation. *Computer Vision and Image Understanding*, 210:103229, 2021.

[336] Can Peng, Kun Zhao, Tianren Wang, Meng Li, and Brian C Lovell. Few-shot class-incremental learning from an open-set perspective. In *European Conference on Computer Vision*, pages 382–397. Springer, 2022.

[337] Liangzu Peng, Paris Giampouras, and René Vidal. The ideal continual learner: An agent that never forgets. In *International Conference on Machine Learning*, pages 27585–27610. PMLR, 2023.

[338] Anastasia Pentina and Christoph Lampert. A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pages 991–999. PMLR, 2014.

[339] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[340] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13846–13855, 2020.

[341] Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. Fetril: Feature translation for exemplar-free class-incremental learning. *arXiv preprint arXiv:2211.13131*, 2022.

[342] Benedikt Pfülb, Alexander Gepperth, and Benedikt Bagus. Continual learning with fully probabilistic models. *arXiv preprint arXiv:2104.09240*, 2021.

[343] Quang Pham, Chenghao Liu, and Steven Hoi. Dualnet: Continual learning, fast and slow. *Advances in Neural Information Processing Systems*, 34:16131–16144, 2021.

[344] Quang Pham, Chenghao Liu, Doyen Sahoo, and HOI Steven. Contextual transformation networks for online continual learning. In *International Conference on Learning Representations*, 2020.

[345] Quang Pham, Chenghao Liu, and HOI Steven. Continual normalization: Rethinking batch normalization for online continual learning. In *International Conference on Learning Representations*, 2021.

[346] Julien Pourcel, Ngoc-Son Vu, and Robert M French. Online task-free continual learning with dynamic sparse distributed memory. In *European Conference on Computer Vision*, pages 739–756. Springer, 2022.

[347] Mozhgan PourKeshavarzi, Guoying Zhao, and Mohammad Sabokrou. Looking back on learned experiences for class/task incremental learning. In *International Conference on Learning Representations*, 2021.

[348] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, pages 524–540. Springer, 2020.

[349] Senthil Purushwalkam, Pedro Morgado, and Abhinav Gupta. The challenges of continuous self-supervised learning. *arXiv preprint arXiv:2203.12710*, 2022.

[350] Chengwei Qin and Shafiq Joty. Lfpt5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5. In *International Conference on Learning Representations*, 2021.

[351] Chengwei Qin and Shafiq Joty. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2776–2789, 2022.

[352] Qi Qin, Wenpeng Hu, Han Peng, Dongyan Zhao, and Bing Liu. Bns: Building network structures dynamically for continual learning. *Advances in Neural Information Processing Systems*, 34:20608–20620, 2021.

[353] Yujia Qin, Jiajie Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Elle: Efficient lifelong pre-training for emerging data. *arXiv preprint arXiv:2203.06311*, 2022.

[354] Haoxuan Qu, Hossein Rahmani, Li Xu, Bryan Williams, and Jun Liu. Recent advances of continual learning in computer vision: An overview. *arXiv preprint arXiv:2109.11369*, 2021.

[355] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[356] Jathushan Rajasegaran, Munawar Hayat, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Random path selection for incremental learning. *Advances in Neural Information Processing Systems*, 3, 2019.

[357] Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. itaml: An incremental task-agnostic meta-learning approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13588–13597, 2020.

[358] Kandan Ramakrishnan, Rameswar Panda, Quanfu Fan, John Henning, Aude Oliva, and Rogerio Feris. Relationship matters: Relation guided knowledge transfer for incremental learning of object detectors. In *CVPR Workshops*, pages 250–251, 2020.

[359] Vinay V Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. *arXiv preprint arXiv:2007.07400*, 2020.

[360] Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*, 2021.

[361] Rahul Ramesh and Pratik Chaudhari. Model zoo: A growing brain that learns continually. In *International Conference on Learning Representations*, 2021.

[362] Amal Rannen, Rahaf Aljundi, Matthew B Blaschko, and Tinne Tuytelaars. Encoder based lifelong learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1320–1328, 2017.

[363] Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[364] Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. In *International Conference on Learning Representations*, 2023.

[365] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[366] Haopeng Ren, Yi Cai, Xiaofeng Chen, Guohua Wang, and Qing Li. A two-phase prototypical network model for incremental few-shot relation classification. In *Proceedings of the International Conference on Computational Linguistics*, pages 1618–1629, 2020.

[367] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[368] Blake A Richards and Paul W Frankland. The persistence and transience of memory. *Neuron*, 94(6):1071–1084, 2017.

[369] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference.

In *International Conference on Learning Representations*, 2018.

[370] Matthew Riemer, Tim Klinger, Djallel Bouneffouf, and Michele Franceschini. Scalable recollections for continual lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1352–1359, 2019.

[371] Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. *Advances in Neural Information Processing Systems*, 31, 2018.

[372] Ryne Roady, Tyler L Hayes, Hitesh Vaidya, and Christopher Kanan. Stream-51: Streaming classification and novelty detection from videos. In *CVPR Workshops*, pages 228–229, 2020.

[373] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[374] Mohammad Rostami, Soheil Kolouri, Praveen Pilly, and James McClelland. Generative continual concept learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5545–5552, 2020.

[375] Mohammad Rostami, Soheil Kolouri, and Praveen K Pilly. Complementary learning for overcoming catastrophic forgetting using experience replay. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3339–3345, 2019.

[376] Mohammad Rostami, Leonidas Spinoulas, Mohamed Hussein, Joe Mathai, and Wael Abd-Almageed. Detection and continual learning of novel face presentation attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14851–14860, 2021.

[377] Subhankar Roy, Mingxuan Liu, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Class-incremental novel class discovery. In *European Conference on Computer Vision*, pages 317–333. Springer, 2022.

[378] Tim GJ Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, and Yarin Gal. Continual learning via sequential function-space variational inference. In *International Conference on Machine Learning*, pages 18871–18887. PMLR, 2022.

[379] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[380] Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2020.

[381] Jonathan Schwarz, Daniel Altman, Andrew Dudzik, Oriol Vinyals, Yee Whye Teh, and Razvan Pascanu. Towards a natural benchmark for continual learning. In *NeurIPS Workshops*, 2018.

[382] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pages 4528–4537. PMLR, 2018.

[383] Ari Seff, Alex Beatson, Daniel Suo, and Han Liu. Continual learning in generative adversarial nets. *arXiv preprint arXiv:1705.08395*, 2017.

[384] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018.

[385] Chenze Shao and Yang Feng. Overcoming catastrophic forgetting beyond continual learning: Balanced training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2023–2036, 2022.

[386] Yilin Shen, Xiangyu Zeng, and Hongxia Jin. A progressive model to enable continual learning for semantic slot filling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1279–1284, 2019.

[387] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in Neural Information Processing Systems*, 34:6747–6761, 2021.

[388] Guangchen Shi, Yirui Wu, Jun Liu, Shaohua Wan, Wenhai Wang, and Tong Lu. Incremental few-shot semantic segmentation via embedding adaptive-update and hyper-class representation. In *Proceedings of the ACM International Conference on Multimedia*, pages 5547–5556, 2022.

[389] Yujun Shi, Kuangqi Zhou, Jian Liang, Zihang Jiang, Jiashi Feng, Philip HS Torr, Song Bai, and Vincent YF Tan. Mimicking the oracle: An initial phase decorrelation approach for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16722–16731, 2022.

[390] Takashi Shibata, Go Irie, Daiki Ikami, and Yu Mitsuzumi. Learning with selective forgetting. In *IJCAI*, volume 2, page 6, 2021.

[391] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9630–9638, 2021.

[392] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in Neural Information Processing Systems*, 30, 2017.

[393] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3400–3409, 2017.

[394] Hyounguk Shon, Janghyeon Lee, Seung Hwan Kim, and Junmo Kim. Dlcft: Deep linear continual fine-tuning for general incremental learning. In *European Conference on Computer Vision*, pages 513–529. Springer, 2022.

[395] Christian Simon, Masoud Faraki, Yi-Hsuan Tsai, Xiang Yu, Samuel Schulter, Yumin Suh, Mehrtash Harandi, and Manmohan Chandraker. On generalizing beyond domains in cross-domain continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9265–9274, 2022.

[396] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. On learning the geodesic path for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1591–1600, 2021.

[397] Pravendra Singh, Pratik Mazumder, Piyush Rai, and Vinay P Namboodiri. Rectification-based knowledge retention for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15282–15291, 2021.

[398] Pravendra Singh, Vinay Kumar Verma, Pratik Mazumder, Lawrence Carin, and Piyush Rai. Calibrating cnns for lifelong learning. *Advances in Neural Information Processing Systems*, 33:15579–15590, 2020.

[399] James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9374–9384, 2021.

[400] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919, 2023.

[401] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[402] Tejas Srinivasan, Ting-Yun Chang, Leticia Leonor Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. Climb: A continual learning benchmark for vision-and-language tasks. *arXiv preprint arXiv:2206.09059*, 2022.

[403] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[404] Serban Stan and Mohammad Rostami. Unsupervised model adaptation for continual semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2593–2601, 2021.

[405] Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. Lamol: Language modeling for lifelong language learning. In *International Conference on Learning Representations*, 2019.

[406] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Distill and replay for continual language learning. In *Proceedings of the 28th international conference on computational linguistics*, pages 3569–3579, 2020.

[407] Qing Sun, Fan Lyu, Fanhua Shang, Wei Feng, and Liang Wan. Exploring example influence in continual learning. *arXiv preprint arXiv:2209.12241*, 2022.

[408] Shengyang Sun, Daniele Calandriello, Huiyi Hu, Ang Li, and Michalis Titsias. Information-theoretic online memory selection for continual learning. In *International Conference on Learning Representations*, 2021.

[409] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua

Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975, 2020.
[410] Siddharth Swaroop, Cuong V Nguyen, Thang D Bui, and Richard E Turner. Improving and understanding variational continual learning. *arXiv preprint arXiv:1905.02099*, 2019.
[411] Shixiang Tang, Dapeng Chen, Jinguo Zhu, Shijie Yu, and Wanli Ouyang. Layerwise optimization by gradient decomposition for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9634–9643, 2021.
[412] Yu-Ming Tang, Yi-Xing Peng, and Wei-Shi Zheng. Learning to imagine: Diversify memory for incremental learning using unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9549–9558, 2022.
[413] Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. Topology-preserving class-incremental learning. In *European Conference on Computer Vision*, pages 254–270. Springer, 2020.
[414] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12183–12192, 2020.
[415] Chen Tessler, Shahar Givony, Tom Zahavy, Daniel Mankowitz, and Shie Mannor. A deep hierarchical approach to lifelong learning in minecraft. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
[416] Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *International Joint Conference on Neural Networks*, pages 1–10. IEEE, 2020.
[417] Michalis K Titsias, Jonathan Schwarz, Alexander G de G Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. In *International Conference on Learning Representations*, 2019.
[418] Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2022.
[419] Marco Toldo and Mete Ozay. Bring evanescent representations to life in lifelong class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16732–16741, 2022.
[420] Hanna Tseran, Mohammad Emtiyaz Khan, Tatsuya Harada, and Thang D Bui. Natural variational continual learning. In *NeurIPS Workshops*, volume 2, 2018.
[421] Ben Tsuda, Kay M Tye, Hava T Siegelmann, and Terrence J Sejnowski. A modeling framework for adaptive lifelong learning with transfer and savings through gating in the prefrontal cortex. *Proceedings of the National Academy of Sciences*, 117(47):29872–29882, 2020.
[422] Gido M van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11(1):1–14, 2020.
[423] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.
[424] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.
[425] Vaibhav Varshney, Mayur Patidar, Rajat Kumar, Lovekesh Vig, and Gautam Shroff. Prompt augmented generative replay via supervised contrastive learning for lifelong intent detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1113–1127, 2022.
[426] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
[427] Tom Veniat, Ludovic Denoyer, and MarcAurelio Ranzato. Efficient continual learning with modular networks and task-driven priors. In *International Conference on Learning Representations*, 2020.
[428] Eli Verwimp, Matthias De Lange, and Tinne Tuytelaars. Rehearsal revealed: The limits and merits of revisiting samples in continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9385–9394, 2021.
[429] Andrés Villa, Kumail Alhamoud, Victor Escorcia, Fabian Caba, Juan León Alcázar, and Bernard Ghanem. vclimb: A novel video class incremental learning benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19035–19044, 2022.
[430] Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57, 1985.
[431] Johannes von Oswald, Christian Henning, Benjamin F Grewe, and João Sacramento. Continual learning with hypernetworks. In *International Conference on Learning Representations*, 2019.
[432] Scott Waddell. Neural plasticity: Dopamine tunes the mushroom body output network. *Current Biology*, 26(3):R109–R112, 2016.
[433] Chengyu Wang, Haojie Pan, Yuan Liu, Kehan Chen, Minghui Qiu, Wei Zhou, Jun Huang, Haiqing Chen, Wei Lin, and Deng Cai. Mell: Large-scale extensible user intent classification for dialogue systems with meta lifelong learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3649–3659, 2021.
[434] Chen Wang, Yuheng Qiu, Dasong Gao, and Sebastian Scherer. Lifelong graph learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13719–13728, 2022.
[435] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. *arXiv preprint arXiv:2204.04662*, 2022.
[436] Jianren Wang, Xin Wang, Yue Shang-Guan, and Abhinav Gupta. Wanderlust: Online continual object detection in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10829–10838, 2021.
[437] Kai Wang, Joost van de Weijer, and Luis Herranz. Acae-remind for online continual learning with compressed feature replay. *Pattern Recognition Letters*, 150:122–129, 2021.
[438] Liyuan Wang, Bo Lei, Qian Li, Hang Su, Jun Zhu, and Yi Zhong. Triple-memory networks: A brain-inspired method for continual learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5):1925–1934, 2021.
[439] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *arXiv preprint arXiv:2310.07234*, 2023.
[440] Liyuan Wang, Kuo Yang, Chongxuan Li, Lanqing Hong, Zhenguo Li, and Jun Zhu. Ordisco: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5383–5392, 2021.
[441] Liyuan Wang, Mingtian Zhang, Zhongfan Jia, Qian Li, Chenglong Bao, Kaisheng Ma, Jun Zhu, and Yi Zhong. Afec: Active forgetting of negative transfer in continual learning. *Advances in Neural Information Processing Systems*, 34:22379–22391, 2021.
[442] Liyuan Wang, Xingxing Zhang, Qian Li, Mingtian Zhang, Hang Su, Jun Zhu, and Yi Zhong. Incorporating neuro-inspired adaptability for continual learning in artificial intelligence. *Nature Machine Intelligence*, 5(12):1356–1368, 2023.
[443] Liyuan Wang, Xingxing Zhang, Qian Li, Jun Zhu, and Yi Zhong. Coscl: Cooperation of small continual learners is stronger than a big one. In *European Conference on Computer Vision*, pages 254–271. Springer, 2022.
[444] Liyuan Wang, Xingxing Zhang, Kuo Yang, Longhui Yu, Chongxuan Li, Lanqing Hong, Shifeng Zhang, Zhenguo Li, Yi Zhong, and Jun Zhu. Memory replay with data compression for continual learning. In *International Conference on Learning Representations*, 2021.
[445] Qiu-Feng Wang, Xin Geng, Shu-Xia Lin, Shi-Yu Xia, Lei Qi, and Ning Xu. Learngene: From open-world to your learning task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8557–8565, 2022.
[446] Runqi Wang, Yuxiang Bao, Baochang Zhang, Jianzhuang Liu, Wentao Zhu, and Guodong Guo. Anti-retroactive interference for lifelong learning. In *European Conference on Computer Vision*, pages 163–178. Springer, 2022.
[447] Rui Wang, Tong Yu, Handong Zhao, Sungchul Kim, Subrata Mitra, Ruiyi Zhang, and Ricardo Henao. Few-shot class-incremental learning for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–582, 2022.
[448] Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 184–193, 2021.
[449] Weikang Wang, Jiajun Zhang, Qian Li, Mei-Yuh Hwang, Chengqing Zong, and Zhifei Li. Incremental learning from scratch for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3710–3720, 2019.
[450] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts

learning with pre-trained transformers: An occam's razor for domain incremental learning. *arXiv preprint arXiv:2207.12819*, 2022.

[451] Zhen Wang, Liu Liu, Yiqun Duan, Yajing Kong, and Dacheng Tao. Continual learning with lifelong vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 171–181, 2022.

[452] Zhen Wang, Liu Liu, Yiqun Duan, and Dacheng Tao. Continual learning through retrieval and imagination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 8, 2022.

[453] Zhen Wang, Liu Liu, Yajing Kong, Jiaxian Guo, and Dacheng Tao. Online continual learning with contrastive vision transformer. In *European Conference on Computer Vision*, pages 631–650. Springer, 2022.

[454] Zirui Wang, Sanket Vaibhav Mehta, Barnabás Póczos, and Jaime G Carbonell. Efficient meta lifelong-learning with limited memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 535–548, 2020.

[455] Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Tiehang Duan, and Mingchen Gao. Improving task-free continual learning by distributionally robust memory evolution. In *International Conference on Machine Learning*, pages 22985–22998. PMLR, 2022.

[456] Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Donglin Zhan, Tiehang Duan, and Mingchen Gao. Meta-learning with less forgetting on large-scale non-stationary task distributions. In *European Conference on Computer Vision*, pages 221–238. Springer, 2022.

[457] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. *arXiv preprint arXiv:2204.04799*, 2022.

[458] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[459] Maciej Wołczyk, Karol Piczak, Bartosz Wójcik, Lukasz Pustelnik, Paweł Morawiecki, Jacek Tabor, Tomasz Trzcinski, and Przemysław Spurek. Continual learning with guarantees via weight interval constraints. In *International Conference on Machine Learning*, pages 23897–23911. PMLR, 2022.

[460] Maciej Wolczyk, Michal Zajac, Razvan Pascanu, Lukasz Kucinski, and Piotr Milos. Disentangling transfer in continual reinforcement learning. *arXiv preprint arXiv:2209.13900*, 2022.

[461] Maciej Wolczyk, Michal Zajkac, Razvan Pascanu, Lukasz Kucinski, and Piotr Milos. Continual world: A robotic benchmark for continual reinforcement learning. *Advances in Neural Information Processing Systems*, 34:28496–28510, 2021.

[462] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022.

[463] Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, and Ali Farhadi. Supermasks in superposition. *Advances in Neural Information Processing Systems*, 33:15173–15184, 2020.

[464] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. *Advances in Neural Information Processing Systems*, 31, 2018.

[465] Guile Wu, Shaogang Gong, and Pan Li. Striking a balance between stability and plasticity for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1124–1133, 2021.

[466] Tongtong Wu, Xuekai Li, Yuan-Fang Li, Gholamreza Haffari, Guilin Qi, Yujin Zhu, and Guoqiang Xu. Curriculum-meta learning for order-robust continual relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10363–10369, 2021.

[467] Tz-Ying Wu, Gurumurthy Swaminathan, Zhizhong Li, Avinash Ravichandran, Nuno Vasconcelos, Rahul Bhotika, and Stefano Soatto. Class-incremental learning with strong pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2022.

[468] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 374–382, 2019.

[469] Yinjun Wu, Edgar Dobriban, and Susan Davidson. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*, pages 10355–10366. PMLR, 2020.

[470] Ziyang Wu, Christina Baek, Chong You, and Yi Ma. Incremental learning via rate reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1125–1133, 2021.

[471] Congying Xia, Wenpeng Yin, Yihao Feng, and S Yu Philip. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1360, 2021.

[472] Ye Xiang, Ying Fu, Pan Ji, and Hua Huang. Incremental learning using conditional adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6619–6628, 2019.

[473] Jiangwei Xie, Shipeng Yan, and Xuming He. General incremental learning with domain-aware categorical representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14351–14360, 2022.

[474] Yang Xie, Peiyao Hu, Junru Li, Jingwen Chen, Weibin Song, Xiao-Jing Wang, Tianming Yang, Stanislas Dehaene, Shiming Tang, Bin Min, et al. Geometry of sequence working memory in macaque prefrontal cortex. *Science*, 375(6581):632–639, 2022.

[475] Ju Xu and Zhanxing Zhu. Reinforced continual learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[476] Kelvin Xu, Siddharth Verma, Chelsea Finn, and Sergey Levine. Continual learning of control primitives: Skill discovery via reset-games. *Advances in Neural Information Processing Systems*, 33:4999–5010, 2020.

[477] Mengqi Xue, Haofei Zhang, Jie Song, and Mingli Song. Meta-attention for vit-backed continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 150–159, 2022.

[478] Shipeng Yan, Lanqing Hong, Hang Xu, Jianhua Han, Tinne Tuytelaars, Zhenguo Li, and Xuming He. Generative negative text replay for continual vision-language pretraining. In *European Conference on Computer Vision*, pages 22–38. Springer, 2022.

[479] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2021.

[480] Shipeng Yan, Jiale Zhou, Jiangwei Xie, Songyang Zhang, and Xuming He. An em framework for online incremental learning of semantic segmentation. In *Proceedings of the ACM International Conference on Multimedia*, pages 3052–3060, 2021.

[481] Guanglei Yang, Enrico Fini, Dan Xu, Paolo Rota, Mingli Ding, Moin Nabi, Xavier Alameda-Pineda, and Elisa Ricci. Uncertainty-aware contrastive distillation for incremental semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[482] Guang Yang, Feng Pan, and Wen-Biao Gan. Stably maintained dendritic spines are associated with lifelong memories. *Nature*, 462(7275):920–924, 2009.

[483] Fei Ye and Adrian G Bors. Learning latent representations across multiple data domains using lifelong vaegan. In *European Conference on Computer Vision*, pages 777–795. Springer, 2020.

[484] Fei Ye and Adrian G Bors. Task-free continual learning via online discrepancy distance learning. *arXiv preprint arXiv:2210.06579*, 2022.

[485] Jingwen Ye, Yifang Fu, Jie Song, Xingyi Yang, Songhua Liu, Xin Jin, Mingli Song, and Xinchao Wang. Learning with recoverable forgetting. In *European Conference on Computer Vision*, pages 87–103. Springer, 2022.

[486] Haiyan Yin, Ping Li, et al. Mitigating forgetting in online continual learning with neuron calibration. *Advances in Neural Information Processing Systems*, 34:10260–10272, 2021.

[487] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020.

[488] Wenpeng Yin, Jia Li, and Caiming Xiong. Contintin: Continual learning from task instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3062–3072, 2022.

[489] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International Conference on Machine Learning*, pages 12073–12086. PMLR, 2021.

[490] Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang. Scalable and order-robust continual learning with additive parameter decomposition. In *International Conference on Learning Representations*, 2019.

[491] Jaehong Yoon, Divyam Madaan, Eunho Yang, and Sung Ju Hwang. Online coreset selection for rehearsal-based continual learning. In *International Conference on Learning Representations*, 2021.

[492] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018.

[493] Longhui Yu, Tianyang Hu, Lanqing Hong, Zhen Liu, Adrian Weller, and Weiyang Liu. Continual learning by modeling intra-class variation. *arXiv preprint arXiv:2210.05398*, 2022.

[494] Lu Yu, Xialei Liu, and Joost Van de Weijer. Self-training for class-incremental semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[495] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6982–6991, 2020.

[496] Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372, 2019.

[497] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017.

[498] Mengyao Zhai, Lei Chen, Jiawei He, Megha Nawhal, Frederick Tung, and Greg Mori. Piggyback gan: Efficient lifelong learning for image conditioned generation. In *European Conference on Computer Vision*, pages 397–413. Springer, 2020.

[499] Mengyao Zhai, Lei Chen, and Greg Mori. Hyper-lifelonggan: scalable lifelong learning for image conditioned generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2246–2255, 2021.

[500] Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong gan: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2759–2768, 2019.

[501] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12455–12464, 2021.

[502] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7053–7064, 2022.

[503] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. *arXiv preprint arXiv:2303.05118*, 2023.

[504] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[505] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1131–1140, 2020.

[506] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *European Conference on Computer Vision*, pages 698–714. Springer, 2020.

[507] Xuchen Zhang, Qian Li, Lianzhang Wang, Zhong-Jian Liu, and Yi Zhong. Active protection: Learning-activated raf/mapk activity protects labile memory from rac1-independent forgetting. *Neuron*, 98(1):142–155, 2018.

[508] Xikun Zhang, Dongjin Song, and Dacheng Tao. Cglb: Benchmark tasks for continual graph learning. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[509] Xinyu Zhang, Tianfang Zhao, Jiansheng Chen, Yuan Shen, and Xueming Li. Epicker is an exemplar-based continual learning approach for knowledge accumulation in cryoem particle picking. *Nature Communications*, 13(1):1–10, 2022.

[510] Yaqian Zhang, Bernhard Pfahringer, Eibe Frank, Albert Bifet, Nick Jin Sean Lim, and Yunzhe Jia. A simple but strong baseline for online continual learning: Repeated augmented rehearsal. *arXiv preprint arXiv:2209.13917*, 2022.

[511] Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. Continual sequence generation with adaptive compositional modules. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3653–3667, 2022.

[512] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13208–13217, 2020.

[513] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[514] Jianjian Zhao, Xuchen Zhang, Bohan Zhao, Wantong Hu, Tongxin Diao, Liyuan Wang, Yi Zhong, and Qian Li. Genetic dissection of mutual interference between two consecutive learning tasks in drosophila. *Elife*, 12:e83516, 2023.

[515] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained gans for generation with limited data. In *International Conference on Machine Learning*, pages 11340–11351. PMLR, 2020.

[516] Na Zhao and Gim Hee Lee. Static-dynamic co-teaching for class-incremental 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3436–3445, 2022.

[517] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9046–9056, 2022.

[518] Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[519] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Co-transport for class-incremental learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 1645–1654, 2021.

[520] Man Zhou, Jie Xiao, Yifan Chang, Xueyang Fu, Aiping Liu, Jinshan Pan, and Zheng-Jun Zha. Image de-raining via continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4907–4916, 2021.

[521] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[522] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. *Advances in Neural Information Processing Systems*, 34:14306–14318, 2021.

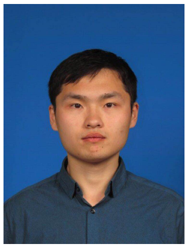
[523] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021.

[524] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6801–6810, 2021.

[525] Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9296–9305, 2022.

[526] Qi Zhu, Bing Li, Fei Mi, Xiaoyan Zhu, and Minlie Huang. Continual prompt tuning for dialog state tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1124–1137, 2022.

[527] Yixiong Zou, Shanghang Zhang, Yuhua Li, and Ruixuan Li. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. *arXiv preprint arXiv:2210.04524*, 2022.



Liyuan Wang is currently a postdoc in Tsinghua University, working with Prof. Jun Zhu at the Department of Computer Science and Technology. Before that, he received the BS and PhD degrees from Tsinghua University. His research interests include continual learning, incremental learning, lifelong learning and brain-inspired AI. His work in continual learning has been published in major conferences and journals in related fields, such as Nature Machine Intelligence, NeurIPS, ICLR, CVPR, ICCV, ECCV, etc.

Xingxing Zhang received the Ph.D. degree in signal and information processing from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 2020, and B.E. degree in 2015. She was also a Visiting Student with the Department of Computer Science, University of Rochester, USA, from 2018 to 2019. She was a Postdoc in the department of computer science, Tsinghua University, Beijing, China, from 2020 to 2022. Her research interests include continual learning, zero/few-shot learning, and data selection. She has received the excellent Ph.D. thesis award from the Chinese Institute of Electronics (CIE) in 2020.

Hang Su , IEEE member, is an associated professor in the department of computer science and technology at Tsinghua University. His research interests lie in the adversarial machine learning and robust computer vision, based on which he has published more than 50 papers including CVPR, ECCV, TMI, etc. He has served as area chair in NeurIPS and the workshop cochair in AAAI22. he received "Young Investigator Award" from MICCAI2012, the "Best Paper Award" in AVSS2012, and "Platinum Best Paper Award" in ICME2018.

Jun Zhu received his BS and PhD degrees from the Department of Computer Science and Technology in Tsinghua University, where he is currently a Bosch AI professor. He was an adjunct faculty and postdoctoral fellow in the Machine Learning Department, Carnegie Mellon University. His research interest is primarily on developing machine learning methods to understand scientific and engineering data arising from various fields. He regularly serves as senior Area Chairs and Area Chairs at prestigious conferences, including ICML, NeurIPS, ICLR, IJCAI and AAAI. He was selected as "AI's 10 to Watch" by IEEE Intelligent Systems. He is a Fellow of the IEEE and an associate editor-in-chief of IEEE TPAMI.