Continual Learning for Histopathology Image Classification in Class Incremental Learning

Yuanyuan Wu[1] wu3yy@mail.uc.edu
Yu Zhao[1] zhao3y3@ucmail.uc.edu
Jun Bai[1] baiju@ucmail.uc.edu
Anca Ralescu[1] ralescal@ucmail.uc.edu
[1] *Computer Science Department, University of Cincinnati, 2600 Clifton Ave, Cincinnati, 45221, OH, USA*

Editors: Under Review for MIDL 2026

Abstract

Continual Learning (CL) is increasingly important for developing adaptive clinical AI models; however, its application to histopathology remains challenging due to strict privacy constraints, expanding diagnostic categories, and substantial staining variability. In this work, we investigate CL for histopathology image classification under a realistic Class-Incremental Learning (CIL) scenario using the NCT-CRC-HE-100K dataset. We benchmark representative regularization-based, replay-based, architecture-based, and prompt-based CL methods to provide a comprehensive evaluation of existing approaches for digital pathology. Among these, prompt-based CL methods have recently emerged as a promising direction by leveraging a frozen pretrained backbone and lightweight learnable prompts to mitigate Catastrophic Forgetting (CF) without storing exemplars or requiring task identifiers during inference. To understand how these methods perform under practical constraints, we analyze the impact of exemplar-free requirements, limited buffer sizes, and training-time budgets across CL paradigms. We further compare four commonly used normalization strategies and find that dataset-level normalization consistently yields the strongest performance. Our results show that replay-based methods achieve the highest accuracy when sufficient memory and training time are available, while prompt-based methods provide competitive exemplar-free performance, making them a practical option for continual adaptation in privacy-sensitive digital pathology workflows.

Keywords: Continual Learning, Catastrophic Forgetting, Prompt-based CL Methods, Histopathology Images.

1. Introduction

Deep learning has achieved remarkable success in Healthcare, particularly in medical image analysis for tasks such as cancer diagnosis from histopathology (Alom et al., 2019; Kather et al., 2018) or disease detection from radiology images (Causey et al., 2019), and various classification and segmentation tasks across X-ray (Rajpurkar et al., 2017), CT (Causey et al., 2019), MRI (Ronneberger et al., 2015), and ultrasound (Baumgartner et al., 2017). Despite these advances, clinical AI models have largely adopted a static "train-once, deploy-forever" paradigm, in which models can not continuously adapt to expanding diagnostic categories, evolving data distributions caused by new acquisition devices, or the gradual distribution shifts that arise as new patient data are collected over time. In dynamic real world healthcare scenarios, clinical AI models increasingly require the ability to continually integrate new medical information while preserving previously acquired knowledge.

The examination of Histopathology images is widely recognized as the gold standard for making final diagnoses of various human lesions (Taqi et al., 2018). Histopathology images are obtained from the microscopic view that captures tissue sections stained with hematoxylin and eosin (H&E). A standard histopathology slide often comprises several hundred thousand individual cells (Shmatko et al., 2022). However, applying CL in histopathology presents unique challenges. Annotated tissue categories are frequently expanded over time as new clinical knowledge is established, requiring models to incorporate novel classes without forgetting previously learned ones. Histopathology images show substantial staining and appearance variability across patients and acquisition sites, which intensifies CF under incremental updates (Komura and Ishikawa, 2018). Furthermore, strict privacy regulations often restrict long-term storage of patient data, limiting the feasibility of replay-based CL methods in medical environments. These challenges motivate the need to explore the effective CL approaches under the CIL scenario tailored to Histopathology images. Therefore, NCT-CRC-HE-100K, a large (100,000 images), diverse (9 classes), and well-curated dataset, is adopted as our benchmark, since it provides diverse tissue types, realistic staining variability, and wide usage in computational pathology research (Kather et al., 2016).

The conventional CL approaches are commonly grouped into regularization-based, replay-based, architecture-based, and prompt-based categories (Qu et al., 2021; Kumari et al., 2025). The first three CL types represent traditional methods, whereas the last type reflects more recent developments in CL. Regularization-based methods, including LwF (Li and Hoiem, 2017), Online EWC (Schwarz et al., 2018), and SI (Qu et al., 2021), constrain parameter updates to preserve previous knowledge. Replay-based methods, such as ER(Rolnick et al., 2019), DER and DER++(Buzzega et al., 2020), alleviate CF by revisiting stored past samples. Architecture-based methods allocate task-specific model parameters or dynamically expand the model to prevent interference across tasks. In this work, we include DyTox (Douillard et al., 2022) as a representative architecture-based method, as it supports CIL without requiring task identities at inference and offers a more scalable design compared to earlier expansion-based approaches. Traditional CL approaches also often train models from scratch as new tasks arrive, making them highly susceptible to CF. With the rapid progress in large-scale representation learning, pretrained models have recently gained significant attention in CL (Janson et al., 2022; Zhang et al., 2023). Leveraging pretrained representations can reduce dependence on replay buffers or strict regularization, and often enhances both stability on past tasks and plasticity when learning new tasks.

Given these benefits of pretrained representations, prompt-based CL methods naturally build upon this paradigm by keeping the backbone frozen and learning only a small set of prompts. Representative examples including L2P (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), and CODAPrompt (Smith et al., 2023), all of which eliminate the need to store exemplars or require task identity at inference time. Despite avoiding memory buffers, prompt-based methods have demonstrated competitive performance compared to replay-based techniques (Wang et al., 2022b,a; Smith et al., 2023). These properties make prompt-based CL especially suitable for exemplar-free CL in medical scenarios(Kumari et al., 2025), especially in digital pathology deployment, where privacy constraints (Kaissis et al., 2020) prohibit retaining patient data and new diagnostic categories emerge over time.

We evaluate DER++ and DualPrompt using four normalization strategies: ImageNet normalization, dataset-level normalization (Tabibu et al., 2019), per-image normalization (Ulyanov et al., 2016), and Macenko stain normalization (Macenko et al., 2009). Because histopathology images differ substantially in color and distribution from natural images, ImageNet normalization is often suboptimal for H&E data. Our results show that DER++ and DualPrompt perform best with dataset-level normalization. Therefore, we adopt this normalization strategy for all subsequent experiments due to its consistently superior performance across CL methods. Moreover, our experimental results show that larger buffer sizes and longer training epochs enable replay-based methods (ER, DER, and DER++) to achieve the highest overall performance. For example, DER++ attains an average accuracy of 94.77 ± 1.82 and a forgetting of 3.66 ± 1.73 when trained for 50 epochs with a buffer size of 500. In contrast, prompt-based methods such as L2P and DualPrompt attain the highest training efficiency with only a few epochs, as they require updating only a small set of prompt parameters rather than the full model. CODA-Prompt, while also a prompt-based approach, requires more training epochs to reach competitive performance. Among these prompt-based methods, DualPrompt performs the best in our experiments, achieving an average accuracy of 88.97 ± 0.60 and a forgetting of 7.70 ± 1.21 at five training epochs.

Our work provides a comprehensive benchmark of prompt-based CL methods alongside representative traditional CL approaches for histopathology images classification under the CIL scenario. The key contributions are summarized as follows:

- We benchmark representative regularization-based, replay-based, architecture-based, and prompt-based CL methods under a realistic CIL setting on the NCT-CRC-HE-100K histopathology dataset.
- We compare four commonly used normalization strategies and show that dataset-level normalization consistently performs best, highlighting its importance for CL in histopathology images.
- We analyze how buffer size and training epochs affect replay-based methods, and how small epochs impact prompt-based methods, providing practical insights into their behavior under real clinical constraints.

2. Related Work

CL aims to enable models to acquire knowledge from a sequence of new tasks without suffering from CF. CF refers to the degradation in performance on previously learned tasks when a model is updated with new ones (Kirkpatrick et al., 2017). Achieving an effective balance between stability and plasticity is a main challenge in CL, where stability helps preserve prior knowledge and mitigates forgetting, while plasticity ensures that the model can effectively learn new tasks. Among various CL scenarios (Van de Ven and Tolias, 2019) in the computer vision field, the most challenging and realistic one is the CIL. In a CIL setting, the model must continually learn new classes without knowing its task identity during inference time. Importantly, CIL has strong clinical relevance in medical imaging, where new disease categories or tissue types are progressively introduced in real workflows, requiring continuous model updates while maintaining diagnostic performance on prior categories (Derakhshani et al., 2022; Ayromlou et al., 2024). Motivated by this practical scenario, our study focuses on CIL for histopathology image classification.

In recent years, several surveys have provided overviews of CL from both the theoretical foundation and application-oriented perspectives. Qu et al. (2021) presents a taxonomy-oriented survey of CL theory, focusing on fundamental challenges such as CF and stability-plasticity trade-offs. Although it thoroughly discusses evaluation protocols and methodological frameworks, it does not delve into domain-specific applications. Wang et al. (2024) further broadens the scope by covering a wide range of CL techniques across multiple domains, including computer vision, reinforcement learning, and natural language processing. Although highly comprehensive in scope, this review remains theoretical and domain-general, lacking discussion on clinical deployment or regulatory considerations. Kumari et al. (2025) provides a comprehensive overview of CL methods for medical imaging analysis, emphasizing the need for CL in medical settings, where models must adapt to evolving data distributions and new tasks without suffering from CF. While these surveys outline the development of CL methods and highlight the importance of medical applications, they do not include empirical evaluations on real medical datasets.

Wu et al. (2024) examines core CL approaches and examines their suitability for different learning scenarios. This survey uses medical imaging examples to illustrate each learning setting and discusses domain-specific challenges in detail, along with commonly used benchmark datasets. They conducted an evaluation of CL methods on MedMNIST (Yang et al., 2023) and reported their comparative performance. Their findings highlight the advantages of exemplar replay and the emerging potential of prompt-based models as exemplar-free alternatives in privacy-sensitive medical environments. However, this work only focuses on low-resolution (28×28) datasets and does not address the unique challenges of histopathology images, such as staining variability and high-resolution (224×224) tissue structures. Moreover, prior benchmarks do not investigate how normalization strategies influence CL performance, nor do they provide detailed analyses of training-time and memory constraints across CL method families. In contrast, our work conducts a comprehensive empirical evaluation on the large-scale NCT-CRC-HE-100K dataset, examines four normalization strategies, and offers method-specific insights into both replay-based and prompt-based CL approaches under realistic clinical constraints.

Lenga et al. (2020) investigate two regularization-based CL methods, containing EWC (Kirkpatrick et al., 2017) and LwF (Li and Hoiem, 2017), within a domain-incremental learning (DIL) scenario using two Chest X-ray datasets. Their experimental results show that Joint Learning achieves the highest performance, as expected, while LwF provides competitive results due to the relatively small domain shift across datasets. However, the CL methods evaluated in their study represent early and relatively simple baselines, and therefore do not reflect the capabilities of more advanced, state-of-the-art CL approaches considered in our work. Moreover, their experiments focus on DIL, where class labels remain fixed across domains, whereas we study the more challenging CIL setting in which new classes must be learned over time without forgetting previous ones. In addition, our work conducts a deeper empirical analysis relevant to clinical deployment, we compare four normalization strategies to determine how image preprocessing affects CL performance, and we examine how practical resource constraints, such as limited buffer sizes and training-time budgets, impact different families of CL methods. These analyses provide insight into which CL approaches are most suitable for real-world clinical AI environments.

3. Methods

3.1. Prompt-Based Methods

In this section, we introduce a family of prompt-based CL methods, including L2P, DualPrompt, and CondaPrompt. These approaches are motivated by the success of prompt learning in Natural Language Processing (NLP) (Liu et al., 2023). Prompt learning incorporates additional instructional and learnable tokens to provide task-specific guidance for a pre-trained model. During task adaptation, the backbone network remains frozen and operates as a generic feature extractor (Wang et al., 2022a). A set of prompt parameters $P \in \mathbb{R}^{L_p \times D}$, where L_p denotes the prompt length that each prompt has L_p tokens, and D is the embedding dimension of each token. Then, it will be prepended to the input tokens to produce an augmented embedding that is fed into the model for downstream prediction. As lightweight and modular components, prompts encode high-level task information and condition the model's behavior without requiring full fine-tuning of the backbone.

Learning to Prompt (L2P) is a prompt-based CL method that introduces a pool of learnable prompts as external memory while keeping the pre-trained ViT backbone frozen (Wang et al., 2022b). For each input image, L2P uses the pre-trained ViT model on ImageNet21K to obtain an embedding feature representation x_e. Next, a query–key matching mechanism is then employed to dynamically select prompts. Specifically, each prompt P_i is related to a learnable key $k_i \in \mathbb{R}^{D_k}$, forming a prompt pool $(K_i, P_i)_{i=1}^{M}$. The query feature $q(x)$ is extracted from the class token output after feeding the input images to the backbone. Then, the similarity scores between the query and keys are computed using the cosine distance function to select the top-N keys, see $s_i = cos(q(x), k_i)$. Therefore, the selected prompts will be added before the input tokens to form the prompted input sequence, $x_p = [P_{s_1}, ..., P_{s_N}]$ for $1 \leq N \leq M$, which is passed through the model and classifier.

L2P optimizes its prompts, keys, and classifier via a combined objective consisting of a supervised classification loss and a key–query alignment loss:

$$L = L(g_\phi(f_r^{avg}(x_p)), y) + \lambda \sum_{K_\infty} \gamma(q(x), k_{s_i}) \tag{1}$$

In Eq.1, k_{s_i} refers to the selected prompt keys, γ is a similarity-based surrogate loss encouraging keys to be updated closely to query features, so when meeting the similar images, the keys will be chosen, and λ is a weighting factor to balance the classification loss and key learning loss. L2P decouples prompt learning strategy and query strategy to make it effectively learn shared prompts and task-specific prompts. In further, L2P expands task-specific capacity and significantly mitigates CF, achieving a strong stability–plasticity balance without replay.

Dual Prompting for Rehearsal-Free Continual Learning(DualPrompt) extends L2P by introducing two complementary types of prompts to address both generic and task-specific knowledge in CL (Wang et al., 2022a). This method also keeps the pre-trained ViT backbone frozen and uses a prompt pool containing the general prompts that capture transferable knowledge shared across tasks and expert prompts that specialize in task-specific information. Prompts are injected at carefully selected MSA layers via a configurable prompting function.

In more detail, a G-prompt $g \in \mathbb{R}^{L_g \times D}$ is globally shared across all tasks throughout the CL process and attached to selected transformer layers to encode task-invariant representations. In contrast, E-prompt $E = e_t{}_{t=1}^{T}$, where $e_t \in \mathbb{R}^{L_e \times D}$, is associated with each Task-$t$and specializes the model to task-unique characteristics. E-prompt is paired with a learnable key enabling the most relevant expert prompt via cosine-similarity matching during inference when task identity is unknown. This retrieval mechanism is conceptually inspired by the query–key strategy in L2P, but differs in that keys represent task-level semantics rather than individual prompts.

After learning prompts, it is crucial to determine where and how to insert them into the MSA layers, since different transformer layers capture different types of knowledge (Raghu et al., 2021). Therefore, the placement and prompting function directly influence how effectively prompts interact with the backbone representations. G-Prompts and E-Prompts are attached to separate, non-overlapping depth ranges to better align with shared versus task-dependent knowledge. In addition, DualPrompt studies two prompting functions: one is a similar way as L2P to prepend prompts to Q, K, V inputs from MLA layer (Prompt Tuning), the second is to divide prompts into p_K, p_V and prepend only to K, V.

To achieve a balance between stability and plasticity, DualPrompt learns the prompts and classifier head through a combined objective. The overall loss consists of a supervised classification term and a constraint term encouraging effective prompt utilization:

$$L = L(f_\phi(f_g, e_t(x)), y) + \lambda L_{match}(x, k_t), x \in D_t \tag{2}$$

The first term in Eq.2 is the cross-entropy loss, while the second term L_match makes query features closer to the correct task key for future retrieval. Here, k_t is the task key, g is the G-Prompt parameters, e_t is the E-Prompt for Task-t, ϕ is the classification head, λ controls the balance between the two loss terms.

By decoupling shared and specialized knowledge pathways, DualPrompt improves learned prompt allocation and enhances cross-task generalization compared to L2P. This dual-prompt structure supports robust CL without replay, offering a stronger stability–plasticity trade-off under CIL.

COntinual Decomposed Attention-based Prompting(CODA-Prompt) (Smith et al., 2023) is an exemplar-free CL method, similar to L2P and DualPrompt, and is likewise categorized as a Prompt-based CL approach. A key distinction between CODA-Prompt and L2P is that all learnable parameters in CODA-Prompt, including prompt components, keys, and attention vectors, are optimized jointly using the standard classification loss, rather than splitting into two separate loss optimizations. This design enables more effective end-to-end optimization and contributes to improved performance.

Instead of selecting a single prompt from a pool, CODA-Prompt introduces a set of learnable *prompt components* and composes the final prompt as a weighted summation of these components:

$$P = \sum_m \alpha_m P_m \tag{3}$$

In Eq. 3, $P \in \mathbb{R}^{L_p \times D \times M}$ denotes the collection of prompt components, P_m is the $m-th$ component, and M denotes the number of components. The weighting vector α is computed for each input according to the similarity between the input query and component keys, optionally modulated by attention vectors, enabling CODA-Prompt to construct input-adaptive prompts that encode task-relevant knowledge.

To refine its prompt selection process, CODA-Prompt applies a learnable attention vector to the query embedding. This element-wise interaction highlights features relevant to each prompt component. Then, the similarity between this new type of query and key vectors determines the mixing weights α_m

$$\alpha = \gamma(q(x) \odot A, K) = \gamma(q(x) \odot A_1, K_1), ..., \gamma(q(x) \odot A_M, K_M) \tag{4}$$

Furthermore, to reduce interference across tasks, CODA-Prompt adds orthogonal constraints on any matrix of prompts, keys, and attention vectors, the loss function is:

$$\mathcal{L} = \mathcal{L}(f_\phi(f_{\theta,P,K,A}(x)), y) + \lambda(\mathcal{L}_{ortho}(P) + \mathcal{L}_{ortho}(K) + \mathcal{L}_{ortho}(A)) \tag{5}$$

where the first term is the cross-entropy loss, the remaining terms are orthogonality regularizers that encourage prompt-related parameters to span distinct subspaces and thereby reduce representational overlap across tasks. Here, θ denotes the frozen parameters of the pre-trained backbone, and ϕ represents the learnable parameters of the linear classification head for the current task.

3.2. Other Compared Methods

CL has been widely explored in medical image analysis using various paradigms, including regularization-based approaches such as SI, replay-based approaches such as ER, DER, and DER++ (Singh et al., 2023). In addition, architecture-based methods like DyTox represent a recent direction in CL. However, despite their promising results in the general vision domain, DyTox has not yet been evaluated in medical imaging. To provide a more comprehensive and fair comparison with prompt-based approaches, we introduce DyTox into the medical image CL setting for the first time.

Online EWC (Schwarz et al., 2018) is a scalable modification of the EWC (Kirkpatrick et al., 2017), designed to regularize deep learning model parameters without increasing memory overhead. Therefore, it is a regularization-based CL method (Kumari et al., 2025). Unlike in standard EWC, which calculates a separate Fisher information matrix for each task, Online EWC maintains a single consolidated Fisher matrix that is updated after each task. Moreover, it uses a decay factor γ to allow the contribution of older tasks to gradually diminish when new tasks are introduced. This way provides a balance between retaining the stability of important knowledge from previous tasks and maintaining plasticity for learning new ones(Kong et al., 2022), while ensuring memory usage remains constant regardless of the number of tasks.

LwF (Li and Hoiem, 2017) is a knowledge-distillation-based CL method (Qu et al., 2021) that regularizes the model by encouraging its prediction on earlier tasks to remain similar to those soft targets generated by the previous model, while simultaneously optimizing the new task loss. The distillation loss function helps preserve learned knowledge through shared feature representations and a unified network architecture, alleviating the CF issue without relying on replay data.

Synaptic Intelligence (SI) is a regularization-based method (Qu et al., 2021) that mitigates CF by estimating how important each model parameter is for previously learned tasks (Zenke et al., 2017). During training, SI accumulates the contribution of each parameter to loss reduction by integrating the product of parameter updates and their corresponding gradients, following a path-integral formulation. These accumulated scores are normalized to form importance weights, which represent how crucial each parameter is for retaining past knowledge. When learning a new task, SI introduces a quadratic regularization term that penalizes changes to parameters with high importance, while allowing less critical parameters to adapt freely. Unlike Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), SI performs this computation online without requiring additional passes over old data or storage of the Fisher matrix, leading to an efficient and memory-light mechanism for knowledge retention across sequential tasks.

Experience Replay (ER) (Rolnick et al., 2019) is a replay-based CL method (Qu et al., 2021), which maintains a small memory buffer to store raw input images and the corresponding hard labels from previous tasks. During the learning of a new task, the model is trained on current task samples in combination with a small batch of stored samples from past tasks. By periodically revisiting examples from earlier tasks, ER preserves previously acquired knowledge through direct supervision and effectively mitigates the CF issue. ER leverages the Reservoir sampling strategy (Vitter, 1985) to ensure that every training example has an equal probability of being stored in the replay buffer, thereby preventing early-task bias. Under this mechanism, ER typically performs well as long as memory capacity is sufficiently large and replay coverage is adequate.

Dark Experience Replay (DER) (Buzzega et al., 2020) is also a replay CL method (Qu et al., 2021) that extends ER by combining rehearsal with knowledge distillation, using both input samples and their stored logits to further mitigate forgetting. During training on each task, DER maintains a memory buffer of input images and their soft-target logits generated from the old model, and jointly optimizes the standard supervised loss on current data and a knowledge distillation loss on replayed samples that matches the model's predictions to the stored logits of replayed samples. By preserving the prior model's output distribution, rather than only relying on hard labels, DER retains richer learned information from old tasks and achieves more robust knowledge retention. DER++ is an extension of DER, proposed by the same authors (Buzzega et al., 2020), which provides an additional supervised loss on replayed memory images, integrating both hard-label replay and soft-logit distillation. This dual-objective design reinforces past knowledge more effectively and improves overall stability compared to DER. It addresses the issue of highly biased logits resulting from a sudden distribution shift caused by the reservoir sampling method.

Dynamic Token Expansion (DyTox) is an architecture-based CL method, specifically designed for the CIL scenario (Douillard et al., 2022) because the transformer architecture is progressing when a new task is introduced to the model. DyTox separates feature extraction and task-specific adaptation using shared Self Attention Blocks (SAB) and Task Attention Blocks (TAB). When a new task is introduced to the model, Dytox dynamically expands a set of task tokens and assigns a new classifier head, while the rest of the model remains shared. This dynamic expansion enables the model to learn new classes without forgetting previously learned knowledge. Rather than using a class token as in ViT (Dosovitskiy, 2020) along with the patch tokens, DyToX adopts the idea that not using a class token in the beginning from CaiT (Touvron et al., 2021) because it can avoid using the learned weights to optimize two contradictory objectives that extracting features from patches and learning useful information for the learning classifier. This aims to better capture task-specific information and reduce optimization conflicts across tasks.

To alleviate the CF issue and encourage diversity among task tokens, DyTox introduces two auxiliary losses, including the distillation loss to prevent forgetting the information from past tasks and the divergence loss to keep the new task token distinct and expressive. This way can make it find a good balance between plasticity and stability. Although DyTox achieves strong performance on natural image benchmarks, to have a remarkable performance on medical image datasets, its architectural expansion strategy requires a large input resolution and training images. As a result, its scalability to small-patch medical images remains limited, making it important to evaluate its transferability in real clinical domains such as histopathology.

4. Dataset Description

To verify the effectiveness of advanced CL methods on medical images, we utilize the NCT-CRC-HE-100K dataset, which consists of 100,000 non-overlapping H&E-stained image patches of human colorectal cancer and normal tissue (Kather et al., 2018). Each patch is extracted from a whole-slide image and represents a localized tissue region such as tumor epithelium, healthy colon mucosa, connective tissue, immune cell infiltrates, adipose tissue, or background regions. The dataset includes nine histological classes, for example, adipose tissue (ADI), background (BACK), and cancerous colon epithelial tissue (TUM) images. Representative images are shown in the Figure.1. As one of the most widely used histopathology datasets for colorectal cancer research, supporting both tumor detection and multi-class colon tissue classification, and enabling robust training and evaluation of deep neural networks on large-scale, patch-level H&E images (Hamida et al., 2021; Ghosh et al., 2021; Sharkas and Attallah, 2024). All images are provided in TIFF format at a resolution of $224 * 224$ pixels with 8-bit depth.

5. Experiments Design and Configuration

We utilized 10 different CL methods: 3 regularization-based methods: SI, Online EWC, LwF; 3 replay-based methods: ER, DER, DER++; 1 architecture-based method: DyTox; 3 prompt-based methods: l2p, DualPrompt, CODA-Prompt. All implementations are based on the Mammoth framework (Cossu et al., 2021) that is widely used in multiple works (Frascaroli et al., 2024; Panariello et al., 2025; Buzzega et al., 2020; Boschini et al., 2022), which offers unified training pipelines, evaluation protocols, and memory-buffer management for CL research. For Dytox experiments, we used the officially released code by the authors, since Mammoth does not include it. Figure.2 displays the overview of the experimental pipeline used to evaluate CL methods on the NCT-CRC-HE-100K histopathology dataset. This workflow includes preprocessing, task construction, selection of CL strategies (regularization-based, replay-based, architecture-based, and prompt-based), and evaluation using average accuracy and forgetting. We split the images from nine classes into three sequential tasks, each containing three classes. At each training stage, a new task is introduced

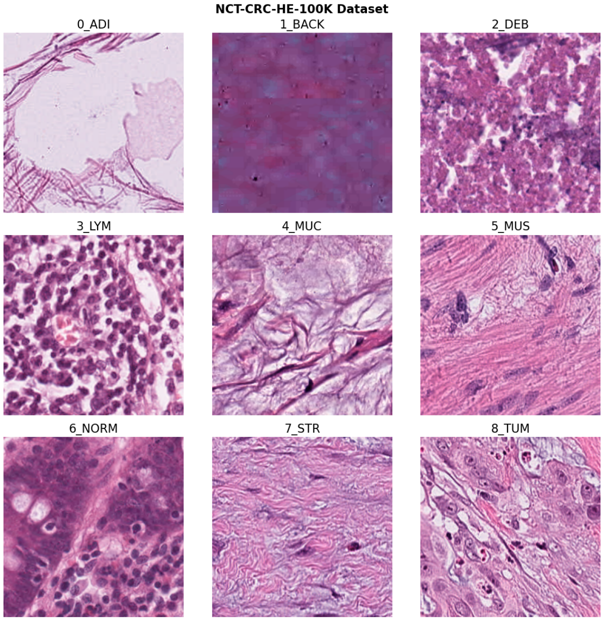

Figure 1: Histopathology Images from NCT-CRC-HE-100K Dataset

into the model, followed by the second and third tasks in a continual fashion. This CIL setup mimics a realistic clinical environment where new diagnostic tasks emerge over time. All methods are trained with a batch size of 64. A summary of hyperparameters is provided in Table 1. All reported results are presented as mean ± standard deviation over three independent runs. All of the data can be found in https://github.com/AILabLLL/CLMedical

In order to study the best performance of these methods, we used their paper suggested parameters and modified them based on the method's characteristics. Firstly, for all of the prompt-based methods, we used the ViT-B/16 model pre-trained on ImageNet-21k and finetuned on ImageNet-1k, as recommended in L2P, DualPrompt, CODA-Prompt. Since the ViT backbone is a pretrained model, only a small number of training epochs is typically required for prompt-based methods. In practice, these methods usually train for 5, 10, or 20 epochs. For example, in the official implementation of L2P, the best performance is obtained with only 5 training epochs. Following this convention, we evaluate the methods within this small epoch range. For the learning rate, usually the prompt-based model utilizes a higher learning rate, such as 0.03 or 0.001 in the released code. That is because prompts are very small modules to be trained while the backbone remains frozen, making prompt optimization stable even under large learning rates. To ensure a fair comparison among all replay-based methods, we adopt ResNet-18 as the backbone architecture trained with the SGD optimizer, following the original design choice of the DER method (Buzzega et al., 2020), which also utilizes ResNet-18 in its experiments.

5.1. Data Normalization Strategies and Pre-processing Steps

Data Normalization is a fundamental pre-processing step in computer vision, as it improves the stability, efficiency, and performance of neural networks. Four commonly used

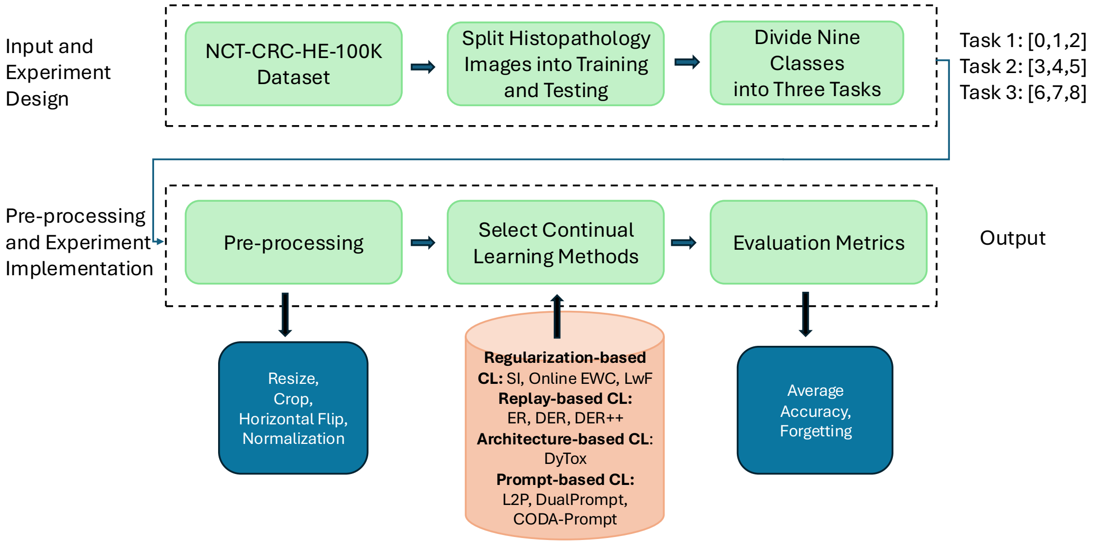
Input and Experiment Design
NCT-CRC-HE-100K Dataset
Split Histopathology Images into Training and Testing
Divide Nine Classes into Three Tasks
Task 1: [0,1,2]
Task 2: [3,4,5]
Task 3: [6,7,8]
Pre-processing and Experiment Implementation
Pre-processing
Select Continual Learning Methods
Evaluation Metrics
Output
Resize, Crop, Horizontal Flip, Normalization
Regularization-based CL: SI, Online EWC, LwF
Replay-based CL: ER, DER, DER++
Architecture-based CL: DyTox
Prompt-based CL: L2P, DualPrompt, CODA-Prompt
Average Accuracy, Forgetting

Figure 2: Experimental Pipeline for Continual Learning on the NCT-CRC-HE-100K Dataset

Table 1: Training hyperparameters for different CL methods

Method	Backbone	Optimizer	Learning Rate
ER/DER/DER++	ResNet-18	SGD	0.03
DyTox	convit	Adam	0.0005
L2P/DualPrompt	ViT-B/16 (frozen)	Adam	0.03
CODA-Prompt	ViT-B/16 (frozen)	Adam	0.001
SI/LwF/EWC	ResNet-18	SGD	0.001

normalization strategies include ImageNet normalization, Dataset-level normalization, Per-image normalization, and Macenko stain normalization. Some common computer vision tasks accept ImageNet normalization values, where

$$\mu = [0.485,\ 0.456,\ 0.406], \quad \sigma = [0.229,\ 0.224,\ 0.225]$$

However, the data distribution of medical images, particularly histopathology images, differs from ImageNet images in color statistics, staining variability, and structural appearance. So we decided to use a dataset-level normalization strategy to calculate the mean and standard deviation from our medical training dataset, ensuring consistent input distribution across samples, where

$$\mu = [0.7408,\ 0.5332,\ 0.7060], \quad \sigma = [0.1645,\ 0.2173,\ 0.1572]$$

Per-image normalization normalizes each image independently using its own mean and standard deviation, reducing brightness and contrast variability across samples. Macenko stain normalization aligns the color appearance of H&E histopathology images by estimating and matching stain vectors in optical density space. To evaluate the effect of different normalization strategies, we compare two representative methods: DualPromp and DER++, under the proposed dataset-level normalization versus other normalization strategies, see experimental results in Section 6.

In addition, we **pre-process** the histopathology images from the NCT-CRC-HE-100K dataset by first resizing them from 224×224 to 256×256 using bicubic interpolation, followed by a random crop back to 224×224 to introduce spatial variability and improve the model's generalization ability. Horizontal flipping is also applied as a standard data-augmentation step. After pre-processing, we compute the dataset-specific mean and standard deviation for each RGB channel using the training images, and use these statistics to normalize the input data.

6. Result Analysis

In the CL field, average accuracy alone is not sufficient to characterize the efficiency of a specific CL method, as it does not explicitly capture the degree of the CF issue. Forgetting (Byun et al., 2023) serves as a complementary metric that is used to quantify how much performance on previously learned tasks degrades after training on new ones. In general, higher final accuracy with lower forgetting represents stronger CL performance.

To assess the effect of normalization strategies on the performance of the CL method, we evaluated two representative methods for DER++ and DualPrompt under four commonly used normalization types. Table 2 demonstrates the corresponding average accuracies. Epochs=50 and Buffer size=500 for DER++, Epochs = 5 for DualPromt. Based on the experimental results, we can observe that the Dataset-level normalization consistently achieves the highest performance for both the DER++ and the DualPrompt methods, indicating that normalization aligned with the dataset's inherent distribution is more effective than ImageNet-based as well as the other two normalization types.

Table 2: Effect of Normalization Type on the Performance of Representative CL Methods

CL Method	Macenko	Per-image	ImageNet Norm	Dataset Norm
DER++	73.95 ± 10.11	92.88 ± 1.83	53.28 ± 37.08	94.77 ± 1.82
DualPrompt	65.72 ± 2.27	79.53 ± 2.52	86.25 ± 1.74	88.97 ± 0.60

For examining how buffer size and number of training epochs influence the effectiveness of replay-based methods, we evaluate ER, DER, and DER++ under combinations of epochs (20, 50) and buffer size (200, 500). As shown in Table 3, DER and DER++ exhibit a clear trend that increasing both the number of epochs and buffer size improves their accuracy while reducing forgetting, with the best performance achieved when $Epochs = 50$ and $Buffer\ size = 500$. ER also benefits from larger buffers and longer training. Overall, DER++ provides the strongest performance when sufficient buffer and training time are available, whereas ER remains competitive under strict resource limitations.

To verify whether prompt-based methods require only a small number of epochs to achieve good generalization, we investigate L2P, CODA-Prompt, and DualPrompt at 5, 10, and 20 epochs. As shown in Table 4, DualPrompt achieves the highest average accuracy and lowest forgetting across the evaluated settings. Meanwhile, both L2P and DualPrompt reach their best performance at 5 epochs, indicating that these methods do not require

Table 3: Effect of Buffer Size and Training Epochs on the Performance of Replay-Based CL Methods

Method	Epochs	Buffer size	Avg. Acc	Forgetting
ER	20	200	68.69 ± 16.83	32.37 ± 13.18
	20	500	73.67 ± 25.95	29.21 ± 27.44
	50	200	87.28 ± 2.81	16.20 ± 7.16
	50	500	**95.07 ± 0.50**	**7.15 ± 0.72**
DER	20	200	48.74 ± 35.34	52.33 ± 42.50
	20	500	92.79 ± 2.38	8.04 ± 4.35
	50	200	82.44 ± 11.94	22.26 ± 17.03
	50	500	**93.99 ± 0.74**	**5.66 ± 2.53**
DER++	20	200	73.52 ± 28.31	28.65 ± 32.10
	20	500	75.70 ± 27.65	26.87 ± 32.04
	50	200	90.85 ± 1.14	9.67 ± 3.25
	50	500	**94.77 ± 1.82**	**3.66 ± 1.73**

extensive training to generalize effectively on histopathology data. This behavior is consistent with the design of L2P and DualPrompt, both of which update a relatively small set of prompt parameters while keeping the backbone frozen, allowing them to converge quickly with only a few training epochs. In contrast, CODA-Prompt employs a different prompt-composition mechanism that benefits from additional training, which explains why its performance continues to improve at 20 epochs. The experimental results suggest that while most prompt-based methods can perform well with limited training, the optimal number of epochs varies across different prompt designs.

On the NCT class-incremental benchmark, regularization methods such as SI, LwF, and EWC still exhibit substantial catastrophic forgetting, even after extensive hyperparameter tuning. Increasing the regularization strengths to favor previously acquired knowledge (e.g., $\alpha = 2.0$ for LwF, $\lambda = 50$ for EWC, and $c = 2.0$ for SI) reduced plasticity but did not prevent collapse, with models showing near-complete forgetting of earlier tasks. These results suggest that, at least on NCT, parameter-space regularization alone is insufficient to address the distribution shift and class interference inherent to CIL.

We additionally evaluate DyTox on the NCT-CRC-HE-100K dataset to provide a more comprehensive comparison. Under 500 training epochs with a memory buffer of 1000, DyTox achieves an average accuracy of 25.09 and a forgetting of 8.59 on the histopathology classification tasks. These results indicate that DyTox does not perform effectively on this dataset in our experimental setting. Furthermore, by jointly examining accuracy and forgetting, we observe that a low forgetting value alone does not necessarily imply strong overall performance. A reliable CL method should exhibit both low forgetting and high average accuracy, as these two metrics together provide a more complete assessment of continual learning effectiveness.

When ample training time and a sufficiently large buffer are available, replay-based methods tend to offer superior performance. In contrast, when rapid convergence is needed or the use of a memory buffer is impractical, prompt-based methods provide a more suitable alternative.

Table 4: Effect of Training Epochs on the Performance of Prompt-based CL Methods

Method	**Epoch**	**Avg. Acc**	**Forgetting**
L2P	5	**83.38 ± 1.39**	**11.42 ± 1.97**
	10	79.67 ± 4.21	19.95 ± 5.30
	20	82.92 ± 2.51	14.80 ± 3.47
CODA-Prompt	5	82.28 ± 2.63	12.40 ± 3.52
	10	82.04 ± 3.26	14.89 ± 5.47
	20	**84.35 ± 2.75**	**9.82 ± 1.51**
DualPrompt	5	**88.97 ± 0.60**	**7.70 ± 1.21**
	10	88.50 ± 1.14	9.46 ± 1.27
	20	87.01 ± 1.33	7.95 ± 2.62

7. Discussion

According to all of the experiments, we can observe that regularization-based methods demonstrate noticeably lower performance compared to replay-based and prompt-based techniques. This is consistent with many existing papers that traditional regularization approaches often struggle on modern CL benchmarks, implying that they may be less suitable for even more complex medical imaging tasks.

In addition, our experiments also emphasize that the normalization strategy plays a substantial role in CL application for histopathology images. In particular, dataset-level normalization consistently outperforms the other normalization methods. This suggests that aligning data preprocessing with the underlying distribution of medical images is important for achieving stable and robust CL.

Ultimately, replay-based methods can achieve strong upper-bound performance when memory resources, training time are not constrained, and when accessing past samples is permitted in real-world applications. However, in realistic clinical environments, such assumptions are often impractical, as mentioned earlier, privacy restrictions frequently limit the storage of patient data, making it impossible to maintain replay buffers. In such situations, exemplar-free CL strategies become very crucial, for instance, prompt-based CL methods. They only train a small set of prompt parameters to guide the model to keep adapting to new tasks without forgetting previously learned knowledge using the frozen backbone and fine-tuning it.

8. Conclusion

In our study, the experimental results indicate that prompt-based CL methods provide a practical exemplar-free alternative to replay-based approaches for clinical deployment, particularly in histopathology CIL settings where patient privacy must be preserved, and data arrive sequentially. In addition, we show that normalization choice is crucial for achieving stable and robust CL performance on medical images, with dataset-level normalization offering the most consistent improvements across all evaluated methods.

References

Md Zahangir Alom et al. Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network. *Journal of Digital Imaging*, 2019.

Sana Ayromlou, Teresa Tsang, Purang Abolmaesumi, and Xiaoxiao Li. Ccsi: Continual class-specific impression for data-free class incremental learning. *Medical Image Analysis*, 97:103239, 2024.

Christian F. Baumgartner et al. Sononet: real-time detection of standard scan planes in fetal ultrasound. In *MICCAI*, 2017.

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.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15920–15930. Curran Associates, Inc., 2020.

Yewon Byun, Saurabh Garg, Sanket Vaibhav Mehta, Praveer Singh, Jayashree Kalpathy-Cramer, Bryan Wilder, and Zachary Chase Lipton. Conditional diffusion replay for continual learning in medical settings. 2023.

Jason L Causey, Yuanfang Guan, Wei Dong, Karl Walker, Jake A Qualls, Fred Prior, and Xiuzhen Huang. Lung cancer screening with low-dose ct scans using a deep learning approach. *arXiv preprint arXiv:1906.00240*, 2019.

Andrea Cossu, Lorenzo Pellegrini, and et al. Mammoth: An extendable, open-source continual learning framework. https://github.com/aimagelab/mammoth, 2021.

Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Tom van Sonsbeek, Xiantong Zhen, Dwarikanath Mahapatra, Marcel Worring, and Cees GM Snoek. Lifelonger: A benchmark for continual disease classification. In *International conference on medical image computing and computer-assisted intervention*, pages 314–324. Springer, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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.

Emanuele Frascaroli, Aniello Panariello, Pietro Buzzega, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Clip with generative latent replay: a strong baseline for incremental learning. *arXiv preprint arXiv:2407.15793*, 2024.

Sourodip Ghosh, Ahana Bandyopadhyay, Shreya Sahay, Richik Ghosh, Ishita Kundu, and KC Santosh. Colorectal histology tumor detection using ensemble deep neural network. *Engineering Applications of Artificial Intelligence*, 100:104202, 2021.

A Ben Hamida, Maxime Devanne, Jonathan Weber, Caroline Truntzer, Valentin Derangère, François Ghiringhelli, Germain Forestier, and Cédric Wemmert. Deep learning for colon cancer histopathological images analysis. *Computers in Biology and Medicine*, 136: 104730, 2021.

Paul Janson, Wenxuan Zhang, Rahaf Aljundi, and Mohamed Elhoseiny. A simple baseline that questions the use of pretrained-models in continual learning. *arXiv preprint arXiv:2210.04428*, 2022.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *(No Title)*, 2018.

Jakob Nikolas Kather et al. Multi-classification of colorectal cancer tissue images using deep learning. *Scientific Reports*, 2016.

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.

Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.

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.

Pratibha Kumari, Joohi Chauhan, Afshin Bozorgpour, Boqiang Huang, Reza Azad, and Dorit Merhof. Continual learning in medical image analysis: A comprehensive review of recent advancements and future prospects. *Medical Image Analysis*, page 103730, 2025.

Matthias Lenga, Heinrich Schulz, and Axel Saalbach. Continual learning for domain adaptation in chest x-ray classification. In *Medical Imaging with Deep Learning*, pages 413–423. PMLR, 2020.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023.

Marc Macenko, Marc Niethammer, Joseph S Marron, David Borland, John T Woosley, Xuemei Guan, Fred Schmitt, and Joan Barker. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. IEEE, 2009.

Aniello Panariello, Emanuele Frascaroli, Pietro Buzzega, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Modular embedding recomposition for incremental learning. *arXiv preprint arXiv:2508.16463*, 2025.

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.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

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.

Maha Sharkas and Omneya Attallah. Color-cadx: a deep learning approach for colorectal cancer classification through triple convolutional neural networks and discrete cosine transform. *Scientific Reports*, 14(1):6914, 2024.

Artem Shmatko, Narmin Ghaffari Laleh, Moritz Gerstung, and Jakob Nikolas Kather. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nature cancer*, 3(9):1026–1038, 2022.

Amritpal Singh, Mustafa Burak Gurbuz, Shiva Souhith Gantha, and Prahlad Jasti. Class-incremental continual learning for general purpose healthcare models. *arXiv preprint arXiv:2311.04301*, 2023.

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.

Sairam Tabibu, PK Vinod, and CV Jawahar. Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Scientific reports*, 9 (1):10509, 2019.

Syed Ahmed Taqi, Syed Abdus Sami, Lateef Begum Sami, and Syed Ahmed Zaki. A review of artifacts in histopathology. *Journal of oral and maxillofacial pathology*, 22(2):279, 2018.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

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

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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. In *European conference on computer vision*, pages 631–648. Springer, 2022a.

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, 2022b.

Xinyao Wu, Zhe Xu, and Raymond Kai-yu Tong. Continual learning in medical image analysis: A survey. *Computers in Biology and Medicine*, 182:109206, 2024.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19148–19158, 2023.