This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Continual-Zoo: Leveraging Zoo Models for Continual Classification of Medical Images

Nourhan Bayasi
University of British Columbia
nourhanb@ece.ubc.ca

Ghassan Hamarneh
Simon Fraser University
hamarneh@sfu.ca

Rafeef Garbi
University of British Columbia
rafeef@ece.ubc.ca

Abstract

In medical imaging, leveraging continual learning (CL) is key for models to adapt to new classes and data distributions without forgetting prior knowledge. Existing CL methods often overlook the use of off-the-shelf pretrained models that are equipped with informative and generalizable representations, opting instead to learn from scratch. In this paper, we propose Continual-Zoo, a novel CL paradigm that smartly leverages a zoo of pretrained models for continual medical image classification. For a given task, Continual-Zoo distills pertinent knowledge from the fixed zoo through cross-knowledge and semantic-knowledge attention mechanisms to obtain class prototypes. Since deploying a zoo could lead to scalability issues with a large number of models, we propose a novel prototypical variational autoencoder, pVAE, as a zoo knowledge encoder. During inference, Continual-Zoo utilizes pVAE as a feature extractor that maps images to the same space of class prototypes and returns the class whose prototype has the shortest distance in the latent space. To mitigate forgetting in CL, pVAE leverages the class prototypes to synthesize images from previously learned tasks before adapting to new ones. Experimental results on various clinical benchmarks demonstrate the superiority of Continual-Zoo over SOTA methods in class-incremental, domain-incremental, and domain and class-incremental learning scenarios, distinguishing it from most CL methods.Code is available at here.

1. Introduction

Deep learning (DL) models are rapidly gaining relevance in medical imaging, excelling in computational tasks like segmentation [14, 15], classification [5, 25], and anomaly detection [50] of vital anatomical structures. In some cases, their capabilities surpass even those of human experts [56], making them a central tool in the advancement of using imaging data for diagnosis. However, these models are typically trained in an offline batch setting, i.e., they assume that all the training data are available at once. In the dynamic healthcare industry, clinical imaging technology, diagnostic workflows, and imaging markers of diseases are subject to constant changes that can significantly impact the accuracy and relevance of deep learning models in real-world applications. Therefore, it is crucial for deep learning models to continuously adapt to the ever-evolving environment to remain effective and relevant in clinical practice.

A significant challenge in this context is catastrophic forgetting [39]. It occurs when a model, in the process of learning new tasks (e.g., new classes or domains), overwrites existing parameters with new data, leading to a loss of previously acquired knowledge. To address this problem, continual learning (CL) has emerged as a promising learning strategy that enables models to learn new tasks sequentially while retaining their performance on previously acquired data [45]. However, existing CL methods usually begin by training models from scratch, often neglecting the potential advantages of integrating off-the-shelf pretrained models into their frameworks. These pretrained models have shown remarkable generalization capabilities and can achieve favorable performance across downstream tasks [34, 38].

Despite their proven effectiveness, adapting pretrained models for continual learning in the medical domain presents substantial challenges. First, the diverse range of knowledge encapsulated in different pretrained models varies considerably in its adaptability across the sequentially introduced datasets, classes, or individual samples. Inadequate handling of this diversity can result in negative forward knowledge transfer [65], leading to poorer performance when compared to randomly-initialized networks [41]. The complexity is compounded by the abundance of the publicly available pretrained models, each with its own architecture, training method, and pretrained data. The selection of a suitable pretrained model for a given task is not a trivial undertaking, as the optimal choice for a current task may not be the best decision for a new incoming task. Second, medical image training datasets are often limited in size as data is collected gradually over time. This presents a notable challenge in finetuning large-scale pre-

trained models on small datasets, as it may result in instability [47] and undermine the model's generalizable representations [63]. Third, experimental evidence suggested that directly adding a CL method on top of a pretrained model may not necessarily lead to improved performance [32]. Consequently, *how can we effectively integrate the knowledge from a diverse set of pretrained models, without relying on fine-tuning, to achieve robust representations suitable for continual medical imaging classification?*

In this paper, we propose Continual-Zoo, a novel two-stage CL pipeline developed to extract relevant knowledge from a zoo of off-the-shelf pretrained models for continual medical image classification. For each task in the training sequence, Continual-Zoo applies two attention techniques on the features extracted from the model zoo to obtain well-representative class prototypes (CPs). The first technique, cross-knowledge attention (CKA), utilizes the inter-knowledge from all models' representations, adaptively emphasizing the most crucial ones for the downstream task. The second technique, semantic-knowledge attention (SKA), utilizes prior, semantic information to further enhance the knowledge obtained from CKA. To overcome potential computational challenges of scalability to a large number of pretrained models during inference, we propose a new prototypical variational autoencoder, pVAE, that acts as a zoo knowledge encoder; i.e., it is trained to map input images to the same space of the prototypes, as illustrated in Fig. 1. Thus, at inference, predictions are made using only the pVAE by returning the class whose prototype in the latent space is closest to the encoder's latent representation of a test image. Furthermore, pVAE creates synthetic samples of past tasks, which reduces forgetting as it adapts to new tasks while also maintaining patient privacy. In summary, our contributions are as follows:

- To the best of our knowledge, Continual-Zoo is the first work that utilizes a zoo of pretrained models within a CL framework for medical image classification.
- We enable an efficient and effective way to leverage the vast knowledge in existing pretrained models toward building representative class prototypes through cross-knowledge and semantic-knowledge attention blocks.
- We design a novel variational autoencoder as a zoo knowledge encoder, used to facilitate inference by mapping a novel test image to the same space of prototypes.
- We assess the performance of Continual-Zoo against multiple baselines and CL methods on three different applications under three CL scenarios. We show that our model achieves SOTA results consistently.

2. Related Work

Continual Learning (CL) models are designed to learn with limited resources from sequentially presented tasks without forgetting. Most techniques are based on regularization or knowledge distillation to minimize changes in parameters when learning new tasks [33, 37, 52]. Parameter isolation is another popular approach, where different subsets of the model parameters are dedicated to different tasks [4, 6, 29, 42]. Alternatively, other methods dynamically expand the network for learning each new task [24]. Another body of work generates pseudo exemplars of the training data for each old task [44, 57] or memorizes old samples in a buffer [3, 49]. While the performance of such replay-based methods is usually the best, several drawbacks are associated with them, including the difficulty in properly selecting representative data or the inability to store real data due to privacy concern, especially in the medical field. Our work does not store exemplars but efficiently generates them using the proposed pVAE.

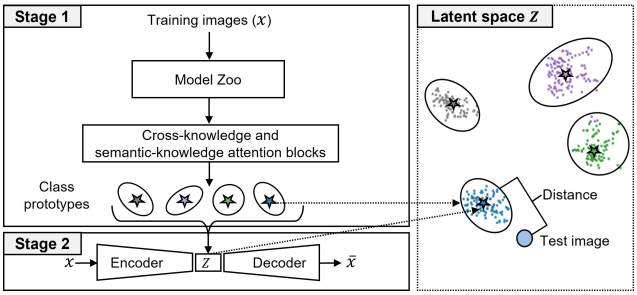

Figure 1. Overview of the proposed framework for continual medical image classification. Continual-Zoo creates class prototypes by leveraging knowledge from a diverse set of pretrained models (model zoo), mining the most beneficial knowledge for a given task through cross-knowledge and semantic knowledge attention mechanisms. These class prototypes are used to regularize the latent space of the proposed pVAE, enabling the mapping of images into the same space as the prototypes. During inference, Continual-Zoo classifies a novel test image by identifying the class prototype closest to the image in the latent space.

Pretrained Models in CL. To the best of our knowledge, there are only a few works in the literature that bridge the gap between CL and pretrained models. L2P [59] and DualPrompt [60] use a pretrained ViT-B/16 network and train a small pool of prompts that update through the CL process. TwF [7] proposes a new strategy that enables a continuous transfer between a source task and incrementally learned tasks. However, L2P, DualPrompt and TwF rely on a 'single' pretrained model and implicitly assumes positive knowledge transfer (i.e., pretrained on ImageNet and evaluated on CIFAR-10). Zhang et al. [66] proposed learnable task-specific adapters within a fixed pretrained model that is used as feature extractor to learn new knowledge of diseases. Recently, CPs generated from large pretrained models have emerged as a simple yet effective solution in CL [28, 40, 46]. This is attributed to the fact that when a model used for extracting representations remains un-

changed, the CPs accumulated for all tasks remain identical, regardless of the number or order of tasks. Additionally, CPs offer a significant advantage in terms of memory cost, as they incur a lower memory footprint compared to using a memory buffer of samples. In contrast to existing CP methods, Continual-Zoo stands out by obtaining CPs from a zoo of diverse pretrained models, rather than relying on a single one, opening up new possibilities for leveraging a wealth of knowledge from multiple models.

CL for Medical Imaging. Prior works explored the applicability of CL methods to medical imaging, mostly adapting existing regularization and generative-replay strategies to the application at hand (e.g., [36, 43, 64]). Recently, Wu et al. [61] proposed a feature-level knowledge distillation technique with contrastive learning to maintain previously acquired knowledge for continual nuclei segmentation. Roy et al. [51] proposed another distillation strategy on mixed-curvature space of the embedding vectors to preserve the complex geometric structure of medical images. Chen et al. [9] proposed a generative replay to substitute images from old tasks with synthetic images. González et al. [17] introduced Lifelong nnU-Net, a nnU-Net based framework for continual training and evaluation of segmentation models in the medical field. Our work is different as it utilizes the knowledge from a zoo of off-the-shelf models for continual medical image classification.

3. Continual-Zoo

We address the problem of leveraging relevant knowledge from a zoo of available pretrained models, without fine-tuning, for continual learning of medical images. The pipeline of Continual-Zoo comprises two stages, demonstrated in Fig. 2. In the first stage, given a zoo of pretrained models (Section 3.2.1), Continual-Zoo applies the proposed CKA and SKA attention techniques (Section 3.2.2, 3.2.3) on the extracted features from the zoo to derive class-specific knowledge, resulting in the formation of well-representative class prototypes (CPs). These CPs are then used in the second stage to regularize a novel prototypical VAE (pVAE) (Section 3.3), which acts as a zoo knowledge encoder to facilitate the inference and mitigate forgetting in CL. At inference, we adopt a task-agnostic inference; predicting the corresponding target from all classes learned so far regardless of the task identity (t) of a given test image (Section 3.4). This differs from typical methods for Continual Learning (CL), where the assumption is that the test input contains a pair (x_{test}, t). Yet, the task identity is not always available in real-world environments.

3.1. Preliminaries

Continual-Zoo learns T tasks sequentially, one at a time, where T is not pre-determined. The t-th task, comprises N_t pairs of input samples $x_t \in \mathcal{X}$ and their corresponding label $y_t \in \mathcal{C}_t$. We use c to denote any class in $\mathcal{C}_t$. During the learning phase of the t-th task, Continual-Zoo does not have access to old data. Continual-Zoo is, to the best of our knowledge, the first comprehensive framework to address three key scenarios in continual learning: (i) Class-incremental learning (CIL), wherein a new task t' involves new, not previously encountered classes, i.e., $\mathcal{C}_{t'}$ s.t. $\mathcal{C}_{t'} \cap \mathcal{C}_t = \phi$; (ii) domain-incremental learning (DIL), wherein a data distribution gap is witnessed across tasks (e.g., changes in scanner manufacturer) but $\mathcal{C}_{t'} = \mathcal{C}_t$; as well as (iii) both domain- and class-incremental learning (DCIL), which is the most challenging scenario.

3.2. Stage 1: Attention-based Class Prototypes

3.2.1 Pretrained Models as Feature Extractors

We start with a pretrained model zoo $\mathcal{F} = \{f^i\}_{i=1}^N$ with N pretrained models used as generic feature extractors. Given training data (x_c, y_c) of a class c from a particular task, as shown in Fig 2 (Stage 1- Model zoo), we extract a set of generic representations $\mathcal{E}_c = \{e_c^i \in \mathbb{R}^{1\times d} \mid e_c^i = f^i(x_c), i = 1, ..., N\}$, which are used to construct a relevant and well-representative CP, referred to as p_c, from the outputs of CKA and SKA, as explained next. For notational clarity, we omit the subscript c hereinafter, as the process is the same for all classes in all tasks.

3.2.2 Cross-Knowledge Attention (CKA)

We propose a multi-head cross-knowledge attention mechanism, inspired by [58], to enable inter-model complementary knowledge fusion. Concretely, we squeeze out the relevant knowledge, denoted as b^i, from model f^i by using the corresponding e^i as the query component, while the rest of the representations in $\mathcal{E}$ are used as the key and value components (Fig. 2; Stage 1- CKA), enabling a more querying knowledge from each model separately. Following [58], the non-projected head of CKA is given in Eq. 1;

$$Att^i = \sum_{j=1, j\neq i}^{N} softmax(\frac{q(e^i)\, k(e^j)^T}{\sqrt{d}})\, v(e^j) \quad (1)$$

where $q(e^i) = e^i W_q^i$, $k(e^j) = e^j W_k^j$, $v(e^j) = e^j W_v^j$, and $W_q, W_k, W_v \in \mathbb{R}^{d\times d}$ are the learnable query, key and value matrices, respectively. The relevant knowledge b^i is;

$$b^i = e^i + concat\,[\, Att_1^i, ..., Att_h^i \,]\, W_o \quad (2)$$

where h is the number of heads and $W_o \in \mathbb{R}^{d\times d}$ is a learnable linear transformation matrix. We repeat this process for N models to extract an averaged class-specific knowledge;

$$g = \frac{1}{N}\sum_{i=1}^{N} b^i. \quad (3)$$

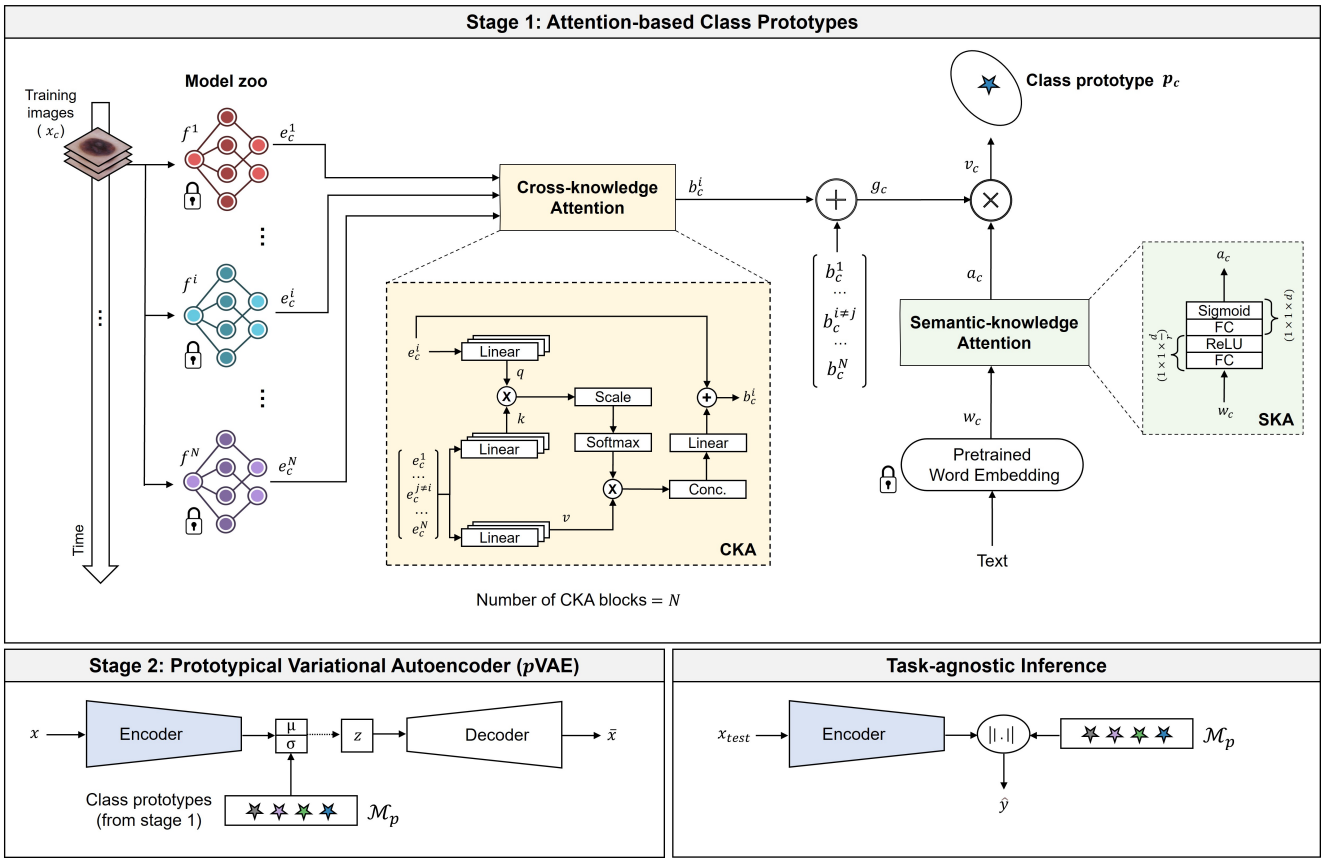

Figure 2. Overview of the Continual-Zoo for the task of continual medical image classification. (Top) In learning stage 1, class-specific knowledge related to each class is extracted from a model zoo and processed to form a prototype for each class using two attention mechanisms: cross-knowledge attention (CKA) and semantic-knowledge attention (SKA), highlighted in yellow and green, respectively. (Bottom-left) In learning stage 2, the class prototypes (CPs) obtained from stage 1 are used to regularize a novel variational autoencoder (pVAE), which facilitates inference and mitigates forgetting in CL. (Bottom-right) During inference, the class with the shortest Mahalanobis distance between its prototype and the test image encoder's distribution is returned for classification.

3.2.3 Semantic-Knowledge Attention (SKA)

Medical images, specifically for skin lesion analysis, presents a high intra-class heterogeneity with respect to the diagnosis. The characteristics of each skin lesion class can significantly change across tasks (Fig.3), requiring the model to learn novel styles of known classes over time. Therefore, inspired by [16], we propose a task-aware encoder, which incorporates prior knowledge derived from the word embeddings of task labels to further enhance the attention on g. To extract this prior knowledge from task labels, we use a fixed pretrained word embedding model as our semantic prior source to obtain an embedding $w \in \mathbb{R}^l$. Then, we feed w into the SKA block to obtain an attention vector $a \in \mathbb{R}^d$. Our SKA is a simple MLP that is equivalent to the excitation module in the SENet [23]; $a = sigmoid\ (FC_d(ReLU(FC_{\frac{d}{r}}(w))))$, where FC_d and $FC_{\frac{d}{r}}$ are fully connected layers with d and $\frac{d}{r}$ neurons, respectively, and r is a parameter choice. The attention vector

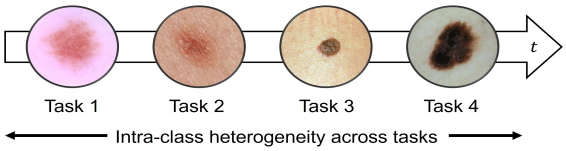

Figure 3. Example of intra-class heterogeneity: The same skin lesion class (e.g., melanocytic nevus) exhibits different appearances across a sequebce of tasks in continual learning. Each task represents a different skin lesion image dataset.

a can be seen as a feature recalibration to obtain $v = a \bigodot g$, where $\bigodot$ is element-wise product operation (Fig. 2; Stage 1- SKA). Finally, we form the d-dimensional class prototype, p, as a Gaussian function modelling the distribution of v over the images of the class, i.e; $p \sim \mathcal{N}(\mu, \sigma)$, where p is parameterized by the mean μ and variance σ, which are stored in memory $\mathcal{M}_p$, with negligible storage. In the

VAE realm, μ and σ will be the parameters of the posterior distribution.

3.3. Stage 2: Prototypical Variational Autoencoder

The continuous updating of the CKA and SKA modules with every incoming data leads to catastrophic forgetting. As we do not store the parameters of these modules, deciding which model parameters to use during inference becomes uncertain. Furthermore, given Continual-Zoo's capability to accommodate an unlimited number of pretrained models, deploying it may introduce scalability concerns, particularly with a large number of pretrained models. To address these challenges, we introduce a novel component: prototypical variational autoencoder, or pVAE. Conceptually, the pVAE acts as a zoo knowledge encoder, enabling a direct mapping of input images into the space of class prototypes (CPs) acquired during the first stage. To achieve this, we propose to regularize pVAE using the distributions of CPs, rather than relying on the conventional VAE's use of a standard normal distribution (Fig. 2; Stage 2).

Additionally, pVAE serves as a crucial component to address the challenge of forgetting in CL and eliminates the need for a replay buffer. Initially, pVAE is trained with images from the first task only. For subsequent tasks, we use the pVAE decoder from the recent learned task to generate synthetic examples, often referred to as pseudo-examples, from the CPs stored in $\mathcal{M}_p$. These pseudo-examples, combined with the training data of the new task, are used to update the projection function such that images of any given class (old or new) are mapped into their corresponding CPs. This way, pVAE effectively retains knowledge from previously encountered tasks, preventing the undesirable loss of information, and ensuring the continual adaptation and learning as the model encounters new data.

3.4. Learning Pipeline and Inference

Optimization of Stage 1. All pretrained models, including the model zoo and the word embedding model, remain fixed throughout the learning process, i.e., only the CKA and SKA modules are optimized to construct the class prototypes of each task. As in other classification problems, the optimization process is facilitated using a classical cross-entropy (CE); $\mathcal{L}_{\text{pred}} = \frac{1}{N_t}\sum_{i=1}^{N_t} \text{CE}\left(h\left(\boldsymbol{v}_i\right), \boldsymbol{y}_i\right)$, where h is a trainable linear layer. While the CE loss encourages features of each class to have a higher projection score on the true class-vector compared to the negative classes, it does not explicitly force different class features to be well-separated. Thus, we add the orthogonal projection loss (OPL) [48], weighted by λ, as a regularization term; $\mathcal{L}_{\text{opl}} = \frac{1}{N_t}\sum_{i=1}^{N_t} \text{OPL}\left(\left(\boldsymbol{v}_i\right); \boldsymbol{y}_i\right)$. The OPL loss enforces inter-class separation alongside intra-class clustering of the prototypes in the feature space through orthogonality constraints. To this end, we define Stage 1 hybrid loss; $\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{pred}} + \lambda\, \mathcal{L}_{\text{opl}}$.

Optimization of Stage 2. We use all the prototypes (i.e., from current and old tasks) to regularize the latent space of pVAE. The regularization is expressed via the Kullback-Leibler (KL) divergence between pVAE encoder's distribution $q_\theta(z, x)$ and the distribution of the CPs. The optimization is thus achieved by minimizing $\mathcal{L}_{pVAE}$;

$$\begin{aligned} \mathcal{L}_{pVAE} &= \mathcal{L}_{rec} + \mathcal{L}_{reg} \\ \mathcal{L}_{rec} &= \|x_c - \bar{x}_c\|^2 \\ \mathcal{L}_{reg} &= KL\left(q_\theta\left(z \mid x_c\right), \mathcal{N}\left(\mu_c, \sigma_c\right)\right). \end{aligned} \tag{4}$$

In Eq. 4, $\mathcal{L}_{rec}$ is the image reconstruction loss, $\mathcal{L}_{reg}$ is the KL regularization loss between the encoder distribution and the prototype distribution of class c, and x_c are the training images (real for the current task, pseudo-examples for the old tasks) belonging to class c.

Inference. In our task-agnostic inference, Continual-Zoo computes the Mahalanobis distance between the latent representation of a novel test image, obtained from the pVAE encoder, and all prototypes in $\mathcal{M}_p$. It assigns the image to the target class $\hat{y}$ associated with the prototype with the minimum distance (Fig. 2; Inference).

4. Experiments and Results

4.1. CL Evaluation Framework

Benchmarks. We evaluate Continual-Zoo on three classification tasks: skin lesion classification from dermatoscopy images (SKIN), peripheral blood cell classification from microscopic images (BLOOD), and colon tissue classification from H&E stained histopathology images (COLON).

SKIN: We use diverse publicly available skin lesion image datasets: HAM10000 (HAM) [55], Dermofit (DMF) [2], Derm7pt (D7P) [31], MSK [19] and UDA [19], which consist of 7,470, 1,212, 959, 3,551, and 601 images, respectively. Each dataset contains skin lesion images from different clinical sites and includes a subset of seven classes.

BLOOD: We utilize the PBS-HCB [1] dataset for peripheral blood cell classification. The dataset includes 17,092 images that are categorized into eight classes.

COLON: We adopt the NCT-CRC-HE [30] dataset for colon tissue classification. The dataset includes 107,180 images belonging to nine tissues.

Setup of CL Scenarios. We assess the performance of Continual-Zoo under three CL scenarios. For CIL, we partition a given dataset into T tasks with non-overlapping classes, denoted as CIL ('dataset_name'). For DIL, we create DIL (SKIN) with four skin lesion image datasets (i.e., $T = 4$), each featuring a unique data distribution while sharing the same classes. For DCIL, we establish DCIL (SKIN) with five skin lesion image datasets (i.e., $T = 5$), each with different distribution and potentially overlapping

classes. Further details on datasets and CL setup are given in Appendix 6.

Evaluation Metrics. We use the accuracy (A) and forgetting measures (F). The accuracy $A = \frac{1}{T}\sum_{i=1}^{T} a_{T,i}$, where $a_{t,i}$ is the balanced-accuracy on the test set of task i after training on the first t tasks, measures the classification accuracy of the model at the end of training averaged across all tasks. The forgetting measure $F = \frac{1}{T-1}\sum_{i=1}^{T-1} \max_{k\in\{1,\dots,T-1\}} a_{k,i} - a_{T,i}$ measures the average difference between the maximum accuracy obtained for task t and its final accuracy.

Baselines and Competitors. We investigate the performance of Continual-Zoo by comparing it with three baselines: SINGLE, which trains separate models for different tasks and deploys a specific model for each task during inference; JOINT, which aggregates the data from all tasks as a consolidated dataset to jointly train a single model (aka. multitask learning); and SeqFT, which finetunes a single model on the current task, without any countermeasure to forgetting. We compare Continual-Zoo against several CL competitors, including two regularization-based methods: EWC [33] and LwF [37]; two generative-based method: DGM [44] and BIR [57]; and two replay-based method: iCaRL [49] and RM [3].

4.2. Implementation Details

Model Zoo. We build Zoo-A of pretrained models, which contains six ResNet-50 models with heterogeneous pretrained data and pretraining schemes (Zoo details are given in Appendix 7). In total, Zoo-A is trained on millions of images across a wide range of computer vision and medical imaging tasks. We also experiment with different zoos in ablation study III.

Semantic Prior Source. We use BioSentVec [11], a biomedical word embedding model trained on PubMed and clinical notes from the Medical Information Mart database to generate a 700-dimensional word vector w.

Implementation Details. We optimize stage 1 in Continual-Zoo using an AdamW optimizer with a batch of 25 images for 100 epochs, having early stopping when overfitting. We set the number of heads h in CKA to 4 and r in SKA to 4. We construct the representation bank $\mathcal{E}$ from the last conv layer of ResNet-50 models, and we use a learnable fully connected layer with each pretrained model to map the features into a lower dimension: 512. We set λ in $\mathcal{L}_{hybrid}$ to 0.05. In stage 2, we use a ResNet-18 model as the pVAE encoder, and a transposed convolutional network as its decoder. We optimize it using an AdamW with a learning rate of 1e-5 and batch size of 25 for 250 epochs. For Continual-Zoo and other generative-based methods, we generate 100 images for each past class. For replay-based methods, comparison considered reserving 50-sample and 100-sample per old class settings. To ensure fairness in comparisons, we use the same backbone as in Zoo-A in all competing methods and we train them using an AdamW optimizer for 100 epochs with early stopping. We initialize all the models with ImageNet pretrained weights and subsequently finetuned them in a supervised manner. We run each experiment with the same set of hyperparameters as in Continual-Zoo and we report the average value on three random tasks ordering.

4.3. Results on SKIN

CIL, DIL, DCIL. The qualitative results of the skin lesion benchmarks in different CL settings are presented in Table 1. We observe the following: 1) In CIL, SINGLE (upper bound) significantly outperforms JOINT in accuracy, which is due to the high heterogeneity among skin lesion classes that adversely affects JOINT's performance. SINGLE addresses this by dividing learning into multiple tasks with fewer classes, thereby minimizing the effect of heterogeneity (e.g., two, two, and three classes for $T = 1, 2$ and 3 in CIL (HAM)). In DIL, SINGLE and JOINT show comparable performance, while JOINT marginally exceeds SINGLE in DCIL. 2) SeqFT suffers from the intense performance degradation due to the challenging model forgetting problem. 3) Zoo-A outperforms other CL methods, especially those relying on regularization techniques. For instance, Zoo-A outperforms EWC by 18.31%, 26.63% and 19.54% on CIL (HAM), DIL (SKIN) and DCIL (SKIN), respectively, indicating that regularization-based methods are ineffective for continual learning of skin lesion images as they are extremely prone to forgetting when learning new tasks. Remarkably, Zoo-A also exceeds the performance of replay-based methods, despite not storing old images. In the context of forgetting, Zoo-A exhibits superior performance with minimal forgetting scores. This can be attributed to the robust image synthesis capability of pVAE that leverages the well-representative and generalizable class prototypes.

Ablation Studies. We conduct diverse ablations to understand the effectiveness of Continual-Zoo's components:

I. Impact of Zoo Size. We explore the impact of the zoo size (i.e., number of pretrained models) on Continual-Zoo's performance. We experiment on the DCIL (SKIN) by sequentially adding ImageNet supervised, CT supervised, MoCo, Mask R-CNN, DeepLabV3 and Keypoint R-CNN models into model Zoo-A. As reported in Table 2, Continual-Zoo outperforms most CL methods in DCIL (SKIN) (refer to Table 1), even with a limited zoo size, i.e., one or two models, indicating its ability to extract compatible knowledge from each model. We also observe a consistent performance boost as the zoo expands, demonstrating Continual-Zoo's adaptability and scalability in leveraging an expanding set of pretrained models.

II. Impact of Zoo Diversity. To assess if Zoo-A's effectiveness stems from the diversity of its pretrained models,

Table 1. Performance evaluation of Continual-Zoo and others on skin lesion benchmarks in CIL, DIL and DCIL settings. Cells in green and blue represent the best and second-best results, respectively.

Method	CIL (HAM)		CIL (DMF)		CIL (D7P)		DIL (SKIN)		DCIL (SKIN)	
	A (↑)	F (↓)	A (↑)	F (↓)	A (↑)	F (↓)	A (↑)	F (↓)	A (↑)	F (↓)
Baselines										
SINGLE⋆	88.35±0.92	-	85.01±0.17	-	73.74±0.11	-	75.12±0.51	-	77.15±0.11	-
JOINT	82.13±0.11	-	80.66±0.15	-	68.32±0.21	-	74.08±0.05	-	80.62±0.09	-
SeqFT	51.54±6.05	50.76±14.61	37.21±5.37	55.25±7.15	36.51±4.25	52.18±5.89	43.28±12.68	69.43±9.87	39.88±9.66	76.85±5.22
CL Methods										
EWC	59.84±3.46	32.29±5.83	50.86±2.03	43.72±6.44	42.73±2.86	38.51±3.71	45.63±7.25	68.69±7.95	51.43±9.85	51.47±12.42
LWF	61.22±4.11	33.70±2.82	49.87±3.85	40.69±9.25	40.67±3.94	35.66±5.66	46.11±6.54	68.37±6.23	49.88±7.76	55.13±9.67
DGM	75.97±1.28	19.27±1.40	64.99±1.49	25.47±2.44	61.24±0.94	22.38±1.34	66.40±0.78	18.72±1.87	59.22±1.89	23.25±1.76
BIR	74.39±1.83	17.85±2.80	61.47±1.72	19.18±2.26	62.9±0.19	19.65±1.61	68.17±0.73	14.35±1.94	62.12±1.44	18.63±2.50
iCaRL (50)	70.80±1.95	18.44±1.26	64.32±1.35	20.17±1.52	60.84±1.86	24.78±1.32	67.03±0.24	15.87±1.18	64.80±1.38	12.51±1.86
iCaRL (100)	73.27±3.20	14.97±1.10	68.49±1.45	18.27±1.14	63.72±1.79	19.28±1.35	70.86±1.43	13.45±1.64	69.12±2.01	11.57±1.39
RM (50)	73.61±0.85	16.83±1.63	63.73±1.26	16.73±1.54	63.05±1.53	22.57±1.24	68.33±1.61	15.79±1.33	63.10±2.16	13.36±1.49
RM (100)	76.32±0.59	15.92±1.28	70.14±1.09	15.22±1.51	65.87±0.93	20.17±1.74	71.18±1.64	14.00±1.16	67.38±1.86	11.32±1.95
Proposed Continual-Zoo										
Zoo-A	78.15±0.85	11.09±1.45	72.51±1.34	14.21±1.26	68.04±1.24	17.58±1.78	72.26±0.88	13.52±1.45	70.97±2.32	10.46±1.92

⋆ We report the average results of the different independent models.

Table 2. Impact of zoo size on DCIL (SKIN). Abbreviations indicate the following: IS (ImageNet Supervised), CS (CT Supervised), MC (MoCo), MR (Mask R-CNN), D3 (DeepLabV3), and KR (Keypoint R-CNN).

Zoo-A	**IS**	**IS+CS**	**IS+CS +MC**	**IS+CS +MC+MR**	**IS+CS+MC +MR+D3**	**IS+CS+MC+ MR+D3+KR**
A (↑)	65.69	67.14	68.02	68.23	69.52	70.97
Parameters	1×	2×	3×	4×	5×	6×

Table 3. Impact of zoo diversity on DCIL (SKIN). Abbreviation indicates the following: IS (ImageNet Supervised).

Zoo-A⋆	**IS**	**2×IS**	**3×IS**	**4×IS**	**5×IS**	**6×IS**
A (↑)	65.69	65.76	65.81	65.94	66.15	66.24
Parameters	1×	2×	3×	4×	5×	6×

we conduct an ablation study using Zoo-A⋆, where the diverse models in Zoo-A are replaced with solely ImageNet supervised models. The findings, presented in Table 3, indicate that the performance remains relatively stable despite an increase in zoo size. Furthermore, comparing Zoo-A⋆ with Zoo-A in Table 2 shows a notable difference in performance, despite both zoos having the same size/number of parameters (all based on ResNet-50). This performance contrast highlights that Continual-Zoo's enhanced performance is primarily driven by the rich and varied knowledge aggregated from the diverse zoo, rather than an increase in models/parameters count.

III. Impact of Model Zoo Choice. To test the influence of the choice of the set of pretrained models, we construct four different zoos with a comparable original performance to Zoo-A. Zoo-B contains CNN-based models with heterogeneous backbones, all pretrained on ImageNet in a supervised fashion. Zoo-C contains transformer-based models, with heterogeneous pretraining schemes (e.g., supervised and self-supervised), all pretrained on ImageNet. Both Zoo-D and Zoo-E have heterogeneous models (e.g., CNNs and ViTs) pretrained on ImageNet but the pretraining scheme is supervised in Zoo-D whereas it is self-supervised in Zoo-E (Refer to Appendix 7 for details). The results in Table 4 indicate that Continual-Zoo with Zoo-E achieves the highest performance among all zoos and across all benchmarks and scenarios. These findings are consistent with earlier research [13, 54], suggesting that self-supervised pretraining plays a significant role in mitigating forgetting and improving the quality of learning. Also, by comparing Zoo-C with Zoo-A and Zoo-B, we notice that ViT-based architectures offer better performance compared to CNNs, implying stronger generalization capabilities on the skin data. For all remaining experiments, we use Zoo-E.

IV. Effect of Cross-Knowledge Attention (CKA). To demonstrate the efficiency of the proposed CKA technique, which extracts relevant knowledge from each pretrained model separately, we train Continual-Zoo, denoted as Continual-Zoo$^{\cancel{CKA}}$, with a single multi-head attention block [58] where the query, key, and value components are formed by stacking all the zoo models' embeddings $\mathcal{E}_c$ into one vector. This setting creates a uniform attention mechanism across all models, akin to turning off our model-adaptive attention approach. The results in Table 4 show a considerable decrease in performance, especially in DIL (SKIN). The stacked vector technique underutilizes each model's compatibility with the downstream task, whereas Continual-Zoo optimally queries knowledge from individual models, leading to enhanced performance.

V. Zoo Contribution. To understand each model's impact in Zoo-E, we calculate its contribution, calculated by dividing its normalized CKA attention score by the total scores from all models, for the different classes in CIL (HAM) and tasks in DIL (SKIN), as shown in Fig. 4. Interestingly, these visualizations reveal varying contribution in class and task levels among models, which further emphasize the importance of our CKA technique.

VI. Effect of Semantic-Knowledge Attention (SKA). To study the impact of leveraging SKA, we train Continual-Zoo without SKA, denoted as Continual-Zoo$^{\cancel{SKA}}$. Table 4 demonstrates a performance drop compared to Zoo-E, emphasizing the importance of SKA in guiding the prototypes in the presence of intra-class heterogeneity across tasks.

VII. Comparison Against Nearest Mean Classifier. Works

Table 4. Performance results from ablation studies evaluating Continual-Zoo on skin lesion benchmarks in CIL, DIL and DCIL settings.

Method	CIL (HAM)		CIL (DMF)		CIL (D7P)		DIL (SKIN)		DCIL (SKIN)	
	A (↑)	F (↓)	A (↑)	F (↓)	A (↑)	F (↓)	A (↑)	F (↓)	A (↑)	F (↓)
Ablation study (III): Choice of Model Zoo										
Zoo-B	73.86±1.36	14.25±1.59	68.92±1.54	15.14±1.37	63.11±1.92	16.22±1.11	67.01±1.13	14.67±1.91	62.58±2.86	15.10±1.87
Zoo-C	78.31±1.72	08.31±1.46	73.18±1.20	08.91±1.17	67.34±1.74	09.05±0.96	74.33±0.74	08.44±1.57	71.26±1.61	10.34±1.76
Zoo-D	76.22±1.14	11.10±1.28	69.50±1.26	11.82±1.67	64.21±1.57	12.33±1.59	68.51±1.46	10.51±1.45	63.71±1.34	11.62±1.49
Zoo-E	79.63±1.26	09.34±1.73	74.98±2.13	10.17±1.45	69.17±1.96	11.46±0.95	75.86±0.44	09.75±1.70	72.80±0.93	10.67±1.35
Ablation Study (IV): Continual-Zoo with Stacked Models' Vectors										
Continual-Zoo ~~CKA~~	77.56±1.65	08.45±1.24	72.82±1.54	08.21±1.43	66.31±0.81	11.03±1.54	71.39±1.76	11.76±1.44	70.53±1.47	10.25±1.73
Ablation Study (VI): Continual-Zoo without Semantic Knowledge										
Continual-Zoo ~~SKA~~	75.61±1.30	10.18±1.35	73.11±1.64	09.92±1.46	67.24±1.87	10.38±1.22	69.95±1.83	10.34±1.70	67.12±1.32	11.26±1.69
Ablation Study (VII): Zoo-E with Nearest Mean Classifier (NMC)										
NMC	72.62±1.25	–	69.91±1.30	–	63.73±1.50	–	65.48±1.14	–	62.71±1.27	–

in [28, 46] have shown that an off-the-shelf pretrained feature extractor itself can be strong enough to achieve a competitive or even better continual learning performance on different classification tasks. To assess this approach's effectiveness in skin lesion classification and compare it with our Continual-Zoo, we use the nearest mean classification (NMC) strategy and calculate the mean features of each class in a task, as following; $\mu_c = \frac{1}{|N_c|} \sum_{x \in N_c} e_{stack}(x)$, where e_{stack} and N_c denotes the stacked features from Zoo-E and the set of training images belonging to class c, respectively. Only class mean features are saved in the memory and are used during evaluation. At the test time, a test sample's stacked feature is extracted from Zoo-E, and the predicted class label is taken as the class whose mean feature is the closest (over all the seen classes so far) to the feature of a test sample; $\hat{y} = \underset{c}{\operatorname{argmin}} \|e_{stack}(x) - \mu_c\|$. Our results in Table 4 demonstrate the NMC's inferior performance compared to our method across all benchmarks and settings. This disparity is due to the domain gap between the pretrained ImageNet data and the specialized skin lesion datasets, emphasizing the importance of incorporating the attention mechanisms to enhance the compatibility with the clinical downstream task.

4.4. Results on BLOOD and COLON

CIL. Table 5 reports the average accuracy and forgetting metrics for CIL (BLOOD) and CIL (COLON). We note that Zoo-E demonstrates superior performance compared to other generative-based CL methods (DGM and BIR) with a reduced forgetting score. This suggests Zoo-E's long-term capability to learn additional tasks while maintaining a commendable overall performance compared to others.

Sequential Analysis. In Appendix 8, we report the running average accuracy to show the performance of Continual-Zoo over time.

5. Conclusion

We proposed Continual-Zoo, a new framework for continual medical imaging classification, which utilizes off-the-shelf pretrained models as a source of knowledge. Continual-Zoo adaptively incorporates attention mechanisms to extract relevant class prototypes for a given task. The prototypes are instrumental for regularizing a novel variational autoencoder, which in return maps a test image to the corresponding prototype's latent space for classification. Through extensive experiments and ablation studies, we demonstrate the superior performance of Continual-Zoo on various CL scenarios and medical data classification tasks while uncovering the factors, such as the size and diversity of the model zoo, that could influence its performance. We hope that this research inspires continued investigation into the utilization of pretrained models for continual learning in medical imaging tasks.

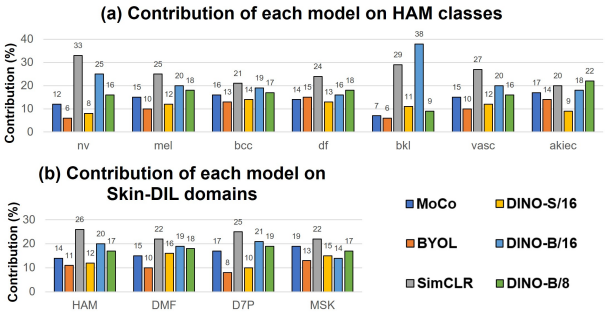

Figure 4. Contribution of each model in Zoo-E on the different HAM classes in CIL (HAM) (a) and tasks in Skin-DIL (b).

Table 5. Performance evaluation of Continual-Zoo and others on four tasks of blood cell and colon tissue benchmarks in CIL setting. Cells in green and blue represent the best and second-best results, respectively, excluding baselines.

Method	CIL (BLOOD)		CIL (COLON)	
	A	F	A	F
SINGLE⋆	97.57±0.10	–	97.80±0.72	–
JOINT	98.13±0.21	–	92.95±1.81	–
SeqFT	33.17±7.65	66.35±14.22	35.16±6.43	76.28± 8.14
EWC	45.73±4.26	63.18±9.67	39.62±4.68	74.48±4.85
LWF	41.84±5.15	65.27±6.38	43.37±7.55	70.59±6.39
DGM	71.43±2.65	24.69±2.34	70.28±1.55	37.21±2.43
BIR	73.78±2.11	23.17±1.95	72.16±1.63	32.96±1.85
iCaRL (50)	78.65±1.67	13.52±2.93	74.22±1.48	10.40±1.72
iCaRL (100)	79.29±1.06	11.43±1.82	76.39±1.29	08.17±1.64
RM (50)	77.46±1.57	17.20±1.95	74.54±1.56	15.39±1.90
RM (100)	78.55±1.44	15.33±1.36	77.67±1.34	13.55±1.52
Zoo-E	80.26±1.88	10.23±1.57	76.58±1.33	09.25±1.42

⋆ We report the average results of four independent models.

References

[1] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30, 2020. 5

[2] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. Springer, 2013. 5

[3] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8218–8227, 2021. 2, 6

[4] Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi. Culprit-Prune-Net: Efficient continual sequential multi-domain learning with application to skin lesion classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 165–175, 2021. 2

[5] Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi. BoosterNet: Improving domain generalization of deep neural nets using culpability-ranked features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 538–548, 2022. 1

[6] Nourhan Bayasi, Siyi Du, Ghassan Hamarneh, and Rafeef Garbi. Continual-GEN: Continual group ensembling for domain-agnostic skin lesion classification. In *ISIC Workshop at International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023. 2

[7] 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. 2

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 1

[9] Boqi Chen, Kevin Thandiackal, Pushpak Pati, and Orcun Goksel. Generative appearance replay for continual unsupervised domain adaptation. *Medical Image Analysis*, 89: 102924, 2023. 3

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[11] Qingyu Chen, Yifan Peng, and Zhiyong Lu. BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5, 2019. 6

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 1

[13] 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. 7

[14] Manju Dabass and Jyoti Dabass. An atrous convolved hybrid seg-net model with residual and attention mechanism for gland detection and segmentation in histopathological images. *Computers in Biology and Medicine*, 155:106690, 2023. 1

[15] Siyi Du, Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi. AViT: Adapting vision transformers for small skin lesion segmentation datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 25–36, 2023. 1

[16] Siyi Du, Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi. Mdvit: multi-domain vision transformer for small medical image segmentation datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 448–458, 2023. 4

[17] Camila González, Amin Ranem, Daniel Pinto dos Santos, Ahmed Othman, and Anirban Mukhopadhyay. Lifelong nnU-Net: a framework for standardized medical continual learning. *Scientific Reports*, 13(1):9381, 2023. 3

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21271–21284, 2020. 1

[19] David Gutman, Noel C. F. Codella, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Nabin K. Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). *arXiv*, abs/1605.01397, 2016. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 1

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 1

[23] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 4, 1

[24] 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*, pages 7797–7805, 2021. 2

[25] Sheng-Kai Huang, Yu-Ting Yu, Chun-Rong Huang, and Hsiu-Chi Cheng. Cross-scale fusion transformer for

histopathological image classification. *IEEE Journal of Biomedical and Health Informatics (JBHI)*, 2023. 1

[26] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. DenseNet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014. 1

[27] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 1

[28] 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. 2, 8

[29] 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 (ICML)*, pages 10734–10750, 2022. 2

[30] Jakob Nikolas et al. Kather. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1), 2019. 5

[31] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *The IEEE Journal of Biomedical and Health Informatics*, 23 (2):538–546, 2018. 5

[32] Zixuan Ke, Bing Liu, Nianzu Ma, Hu Xu, and Lei Shu. Achieving forgetting prevention and knowledge transfer in continual learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:22443–22456, 2021. 2

[33] 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. 2, 6

[34] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2661–2671, 2019. 1

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[36] 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, 2020. 3

[37] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 2, 6

[38] Ziyue Li, Kan Ren, Xinyang Jiang, Bo Li, Haipeng Zhang, and Dongsheng Li. Domain generalization using pretrained models without fine-tuning. *arXiv preprint arXiv:2203.04600*, 2022. 1

[39] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, pages 109–165. Elsevier, 1989. 1

[40] Mark D McDonnell, Dong Gong, Amin Parveneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. *arXiv preprint arXiv:2307.02251*, 2023. 2

[41] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. Factors of influence for transfer learning across diverse appearance domains and task types. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9298–9314, 2021. 1

[42] Zichen Miao, Ze Wang, Wei Chen, and Qiang Qiu. Continual learning with filter atom swapping. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022. 2

[43] Ana C Morgado, Catarina Andrade, Luís F Teixeira, and Maria João M Vasconcelos. Incremental learning for dermatological imaging modality classification. *Journal of Imaging*, 7(9):180, 2021. 3

[44] 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 (CVPR)*, pages 11321–11329, 2019. 2, 6

[45] 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. 1

[46] Francesco Pelosin. Simpler is better: off-the-shelf continual learning through pretrained backbones. In *(CVPR) Workshop on Transformers for Vision*, 2022. 2, 8

[47] Matthew E Peters, Sebastian Ruder, and Noah A Smith. To tune or not to tune? adapting pretrained representations to diverse tasks. In *the 4th Workshop on Representation Learning for NLP (RepL4NLP'19)*, 2019. 2

[48] Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Orthogonal projection loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12333–12343, 2021. 5

[49] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental classifier and representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2001–2010, 2017. 2, 6

[50] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2155–2162, 2023. 1

[51] Kaushik Roy, Peyman Moghadam, and Mehrtash Harandi. L3DMC: Lifelong learning using distillation via mixed-curvature space. *MICCAI*, 2023. 3

[52] Marco Toldo and Mete Ozay. Bring evanescent representations to life in lifelong class incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16732–16741, 2022. 2

[53] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357, 2021. 1

[54] Tuan Truong, Sadegh Mohammadi, and Matthias Lenga. How transferable are self-supervised features in medical image classification tasks? In *Machine Learning for Health*, pages 54–74, 2021. 7

[55] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 5

[56] Philipp Tschandl, Cliff Rosendahl, Bengu Nisa Akay, Giuseppe Argenziano, Andreas Blum, Ralph P Braun, Horacio Cabo, Jean-Yves Gourhant, Jürgen Kreusch, Aimilios Lallas, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA dermatology*, 155(1):58–65, 2019. 1

[57] 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. 2, 6

[58] 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 (NeurIPS)*, 30, 2017. 3, 7

[59] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 139–149, 2021. 2

[60] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[61] Huisi Wu, Zhaoze Wang, Zebin Zhao, Cheng Chen, and Jing Qin. Continual nuclei segmentation via prototype-wise relation distillation and contrastive learning. *IEEE Transactions on Medical Imaging*, 2023. 3

[62] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9981–9990, 2021. 1

[63] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. AIM: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023. 2

[64] Jingyang Zhang, Ran Gu, Peng Xue, Mianxin Liu, Hao Zheng, Yefeng Zheng, Lei Ma, Guotai Wang, and Lixu Gu. S 3 r: Shape and semantics-based selective regularization for explainable continual segmentation across multiple sites. *IEEE Transactions on Medical Imaging*, 2023. 3

[65] Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 2022. 1

[66] Wentao Zhang, Yujun Huang, Tong Zhang, Qingsong Zou, Wei-Shi Zheng, and Ruixuan Wang. Adapter learning in pretrained feature extractor for continual learning of diseases. *MICCAI*, 2023. 2