CLIP MODEL IS AN EFFICIENT CONTINUAL LEARNER

Anonymous authors
Paper under double-blind review

ABSTRACT

The continual learning setting aims to learn new tasks over time without forgetting the previous ones. The literature reports several significant efforts to tackle this problem with limited or no access to previous task data. Among such efforts, typical solutions offer sophisticated techniques involving memory replay, knowledge distillation, model regularization, and dynamic network expansion. The resulting methods have a retraining cost at each learning task, dedicated memory requirements, and setting-specific design choices. In this work, we show that a frozen CLIP (Contrastive Language-Image Pretraining) model offers astounding continual learning performance without any fine-tuning (zero-shot evaluation). We evaluate CLIP under a variety of settings including class-incremental, domain-incremental and task-agnostic incremental learning on five popular benchmarks (ImageNet-100 & 1K, CORe50, CIFAR-100, and TinyImageNet). Without any bells and whistles, the CLIP model outperforms the state-of-the-art continual learning approaches in majority of the settings. We show the effect on CLIP model's performance by varying text inputs with simple prompt templates. To the best of our knowledge, this is the first work to report the CLIP zero-shot performance in a continual setting. We advocate the use of this strong yet embarrassingly simple baseline for future comparisons in the continual learning tasks.

1 INTRODUCTION

Traditionally, deep neural networks (DNNs) trained in a supervised manner on training sets comprising of all the classes of interest have shown excellent results. Such models can presumably learn all the relevant features from the dataset in a single training episode. However, in real world, all data samples may not be available at once. To cater for such scenarios, continual learning provides a promising paradigm, since it enables learning where the data distribution shifts over time. DNNs trained on such incremental stream of data, however, suffer from catastrophic forgetting since the previous task data can not be accessed in its entirety (McCloskey & Cohen, 1989).

In the literature, four popular continual learning protocols exist; Task-incremental learning (TIL) use task-specific neural networks, where the task identity is assumed to be known at the inference. Class-incremental learning (CIL) settings add classes in a sequence with the task identity unknown at the inference. In Domain-incremental learning (DIL), the number of classes remain the same but the data domain evolves over time. Task-free (or task-agnostic) continual learning (TFCL) is a more general setting where there are no explicit task boundaries and data can appear freely in the continual learning phases (De Lange et al., 2021). The major challenge faced by all these methods is to avoid forgetting previously learned knowledge while updating on new data.

Several specialized methods have been developed in continual learning literature to reduce catastrophic forgetting. Among such methods, typical solutions offer sophisticated techniques involving memory replay (Rebuffi et al., 2017; Shin et al., 2017; Lopez-Paz & Ranzato, 2017), knowledge distillation (Hinton et al., 2015; Li & Hoiem, 2017), model regularization (Kirkpatrick et al., 2017), parameter isolation (Mallya & Lazebnik, 2018; Fernando et al., 2017), and dynamic network expansion (Yan et al., 2021; Douillard et al., 2022; 2020). The resulting methods have a retraining cost at each learning task, need dedicated memory for storing exemplars or past models, and involve complex hyper-parameter tuning which limits their practical utility. Furthermore, the above continual learning protocols are generally addressed separately and the existing approaches involve setting-specific design choices making them non-transferable across different continual learning settings.

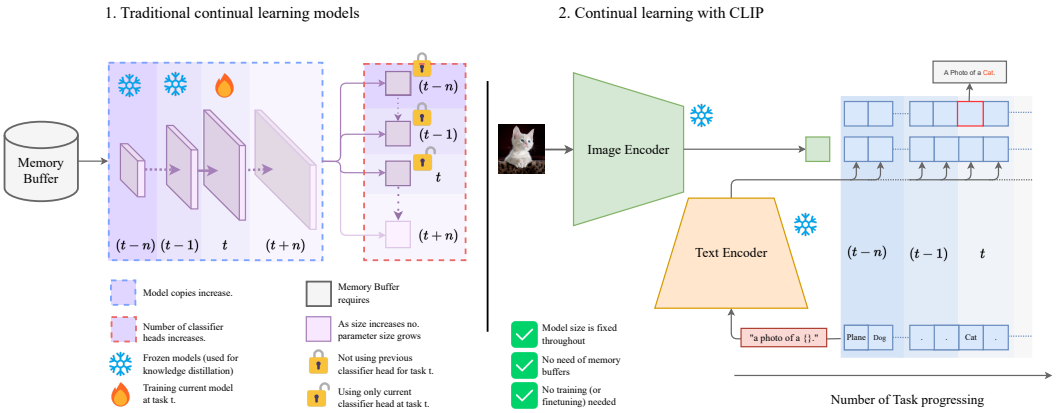

Figure 1: *Left:* Traditional continual learning approaches can require memory buffers, complex hyper-parameter tuning, saving a copy of previous models and their number of classifier heads increases at each learning step. *Right:* Our goal is to design a simplistic unified model that works well across multiple continual learning settings without incurring task-wise training, dedicated memory requirements and careful hyper-parameter selection. A CLIP-based continual model is shown to perform exceptionally well on a number of continual learning settings without requiring any training/fine-tuning, memory buffers or increase in model size with the growing number of tasks.

In this work, we aim to test the progress made so far towards a truly continual learning system. Our main question is to explore if the state-of-the-art narrow models can be replaced with a simple generic approach that does not require training for each incremental step, works without any exemplar memory storage and can work across all of the existing incremental learning protocols with minimal or no hyper-parameter tuning. To this end, we show that a frozen CLIP model (Radford et al., 2021) offers great promise due to its generalizable representations and zero-shot behaviour without requiring any parameter-tuning. Figure 1 gives an overview of traditional continual learning methods and frozen CLIP in the continual learning system.

Our extensive evaluations across four diverse settings (TIL, CIL, DIL, and TFCL) and seven datasets (ImageNet-100 & 1K (Deng et al., 2009), CLEAR (Lin et al., 2021), CIFAR100 (Krizhevsky et al., 2009), TinyImageNet (Stanford), CoRe50 (Lomonaco & Maltoni, 2017) and Gaussian scheduled CIFAR-100 (Shanahan et al., 2021)) demonstrate CLIP's competitiveness on all these CL settings. This generalization behaviour is due to the large-scale pre-training of vision-language model like CLIP that optimizes contrastive training objective on 400M image-text pairs scraped from the internet. During pre-training, CLIP learns a diverse range of high-level representations which are transferable to multiple downstream tasks including the incremental tasks. We also show how simple prompt engineering for text inputs affects the CLIP's performance for CL.

In summary, this work layouts the baseline for the future direction in continual learning based on the pre-trained vision-language models. We evaluate the pre-trained frozen CLIP model in a variety of continual learning settings on popular image recognition benchmarks and compare to current state-of-the-art methods to show that out-of-box CLIP representations perform competitively in all cases. Our results aim to consolidate the fragmented efforts in continual learning landscape that work on specific settings, highlighting the need for generic approaches that can work across multiple settings.

2 Related Works

Continual Learning: The existing continual learning methods mostly employ one of the following schemes: (1) model regularization (2) memory replay, and (3) dynamic network expansion. Model Regularization-based methods (Kirkpatrick et al., 2017; Aljundi et al., 2018; Kirkpatrick et al., 2017; Li & Hoiem, 2017) avoid catastrophic forgetting by limiting the plasticity of the model parameters that are important for previous tasks. Even though these methods do not require memory replay, they have only been shown to work on simplistic task-incremental setting (the task identity is assumed to be known at inference time) and on smaller datasets (Wu et al., 2019). Memory Replay based meth-

ods use exemplars that are either stored in the memory, or synthesized using generative techniques, and are effective on more challenging settings and datasets (Rebuffi et al., 2017; Kamra et al., 2017; Buzzega et al., 2020; Cha et al., 2021). However, their performance degrades for smaller memory size (Cha et al., 2021), and storing these exemplars can introduce security and privacy concerns Shokri & Shmatikov (2015). Architecture-driven CL methods either dynamically expand a network Rusu et al. (2016); Li et al. (2019b); Zhao et al. (2022), or divide into sub-networks to cater for the new tasks (Zhao et al., 2022; Wang et al., 2020; Ke et al., 2020; Rajasegaran et al., 2019a). Such approaches lack scalability, since the network capacity grows with tasks.

Vision-language Models: Training a joint vision-language embedding space enables interaction amongst text and image data, and is critical to solve problems such as zero-shot learning, visual grounding, and image captioning. While the initial vision-language models were single-stream and processed the concatenated input from visual and text data as a single set of input (Li et al., 2019a; Kim et al., 2021), more recent approaches, such as the Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) are dual-stream with dedicated encoders for image and text inputs. The representations from the two encoders are projected into a unified embedding space, and a contrastive learning objective is employed to minimize the distance between matching image-caption pairs, and maximize otherwise. Subsequent works have shown that CLIP is scalable, and its capabilities improve once trained on large-scale noisy data of 1 billion non-curated samples Jia et al. (2021). The representations learned by the CLIP have been shown to generalize well across numerous downstream tasks, including image-text retrieval, with excellent zero-shot transfer capabilities (Li et al., 2021a). CLIP can also be flexibly adapted to videos (Yuan et al., 2021; Xu et al., 2021; Fang et al., 2021), and to capture object-level interactions (Yao et al., 2021; Li et al., 2021a; Rasheed et al., 2022). However, the applicability of CLIP representations for CL is not yet investigated.

With recent advances in Vision Transformers (Khan et al., 2021), and prompt-based fine-tuning in NLP (Li & Liang, 2021), Wang *et al.* have shown that interacting with an ImageNet pre-trained model via prompt learning is a promising approach for continual learning (Wang et al., 2022a;b). A small set of learnable parameters, called prompts, is appended to the input, and enables quick adaptation of a frozen ImageNet pre-trained model to new streaming tasks. In our approach, we show that directly leveraging the pre-trained vision-language model, without introducing any learnable parameters, is a simple yet promising approach to continual learning. We argue that adapting a joint vision-language model like CLIP (Radford et al., 2021) for continual learning presents multiple advantages. It enables catering for practical scenarios where there are no well-defined task identities and boundaries, and the model is required to dynamically adapt to streaming data in a task-agnostic manner. As such, leveraging from the CLIP model requires no compute expensive training or fine-tuning on the data for new tasks. Further, in contrast to the current state-of-the-art methods that require a memory buffer to store training examples from previous tasks, our approach is rehearsal free, and is more suitable for scenarios where storing training examples could have practical privacy and security concerns or storage constraints (Shokri & Shmatikov, 2015). Instead of storing past samples in the memory buffer, where the performance can deteriorate for small buffer size, our approach requires a constant memory throughout all the learning episodes. In summary, our approach is memory free, does not require test-time task identity information, can be flexibly and easily adapted to any number of classes without requiring any additional learnable parameters.

3 METHODOLOGY

In this section, we first briefly discuss different continual learning settings. Next, we introduce Contrastive Language-Image Pre-training (CLIP, Radford et al. (2021)) and explain how to apply it to different downstream continual learning tasks in a zero shot manner with hand-crafted prompts. WE call the CLIP evaluated on a diverse set of continual learning settings as *Continual-CLIP*.

3.1 CONTINUAL LEARNING FORMULATION.

Different Continual Learning (CL) problems focus on training models on a non-stationary data from sequential tasks, while reducing forgetting on the old tasks[1]. Consider a sequence of tasks $D = \{D_1, D_2, \ldots, D_T\}$, where the t^{th} task $D_t = \{(\boldsymbol{x}_i^t, y_i^t)\}_{i=1}^{N_t}$ contains tuples of input samples $\boldsymbol{x}_i^t \in \mathcal{X}$

[1]We use term task for a distinct training episode happening in CL.

and its corresponding label $y_i^t \in \mathcal{Y}$. The goal is to optimize the model $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ parameterized by θ, such that it predicts the label $y = f_\theta(\boldsymbol{x}) \in \mathcal{Y}$ given an unseen test sample $\boldsymbol{x}$ from an arbitrary task. During task t, the data from the previous distribution D_{t-1} is not available or restricted.

Based on the input $P(\mathcal{X}^{(t)})$ and output $P(\mathcal{Y}^{(t)})$ distributions of task t, with $P(\mathcal{X}^{(t)}) \neq P(\mathcal{X}^{(t+1)})$, continual learning can be classified into four popular settings with slightly different assumptions *i.e.*, task-incremental, class-incremental, domain-incremental and task-free CL. The common task-, class-, domain-incremental settings assumes task data D_t arrives in a sequence such that $t = \{1, 2, \ldots, T\}$. At task t, the class-incremental setting defines output space for all observed class labels $\mathcal{Y}^{(t)} \subset \mathcal{Y}^{(t+1)}$ with $P(\mathcal{Y}^{(t)}) \neq P(\mathcal{Y}^{(t+1)})$. Different from class-incremental setting, task-incremental settings defines $\mathcal{Y}^{(t)} \neq \mathcal{Y}^{(t+1)}$ with $P(\mathcal{Y}^{(t)}) \neq P(\mathcal{Y}^{(t+1)})$ requires tasks label t to indicate isolated output heads $\mathcal{Y}^{(t)}$. Different from task- and class-incremental settings where each task has different classes, domain incremental setting is defined as $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t+1)}$ with $P(\mathcal{Y}^{(t)}) = P(\mathcal{Y}^{(t+1)})$ and $P(\mathcal{X}^{(t)}) \neq P(\mathcal{X}^{(t+1)})$ such that it contains a set of images drawn from a different domain, but has the same set of classes for every task. The more challenging setting is task-free (or task-agnostic) setting, where the task data D_t changes smoothly and the task identity t is unknown (De Lange et al., 2021). In this work, we perform extensive experiments to show the effectiveness of our CLIP based approach on the challenging class- and domain-incremental settings, as well as the task-agnostic setting in CL.

3.2 CONTINUAL-CLIP

Contrastive Language-Image Pre-training (CLIP, Radford et al. (2021)) consists of two parallel encoders, one is for text and the other for images. The text encoder is based on the Transformer (Vaswani et al., 2017) architecture which generates embedding representations for the text-based language inputs. On the other hand, the image encoder architecture can be based on a CNN *e.g.*, ResNet-50 He et al. (2016) or a Vision Transformer (ViT) model (Dosovitskiy et al., 2020) to transforms the high-dimensional input images into a compact embedding space. The embedding feature dimension for the text and image encoder are same, thus enabling learning a shared and unified representation space for the two modalities.

The CLIP model is trained with a contrastive loss which promotes the similarities between image and text embeddings belonging to the same image-caption pair, so that both get aligned in the joint feature space. Given a batch of image-text pairs, CLIP objective is to maximize the cosine similarity between matched pairs while minimize the similarity between unmatched image-text embedding pairs. Using this learning objective, the model is trained on a large scale dataset of 400M image-caption pairs, and it learns highly transferable representations for image and text data, which have demonstrated impressive zero-shot generalization capabilities.

Let us denote CLIP model as $\mathcal{F} = \{\mathbf{E}_{\texttt{visual}}, \mathbf{E}_{\texttt{text}}\}$, where $\mathbf{E}_{\texttt{visual}}$ and $\mathbf{E}_{\texttt{text}}$ are image and text encoders respectively. Consider a K-class classification problem, such that a single test image $\boldsymbol{x}_{test} \in \mathbb{R}^{C \times H \times W}$ belongs to a class $y \in \mathbb{R}^K$. In the traditional zero-shot classification scenario, every $y_i \in \mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$ is prepended by a hand-crafted prompt template $\boldsymbol{p}$, such as "$\texttt{a photo of a } \{\texttt{category}\}$" to form a category-specific text input $\{\boldsymbol{p}; y_i\}$. Then this text input is fed to the text-encoder to get the text embedding $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_K\}$, where $\boldsymbol{t}_i = \mathbf{E}_{\texttt{text}}(\{\boldsymbol{p}; y_i\})$. The image input $\boldsymbol{x}_{test}$ is fed to the $\mathbf{E}_{\texttt{visual}}$ to get the corresponding image embedding $\boldsymbol{v} = \mathbf{E}_{\texttt{visual}}(\boldsymbol{x}_{test})$. Both the text and image embeddings are then matched to compute the similarity score $\boldsymbol{s}_i = \texttt{sim}(\boldsymbol{t}_i \cdot \boldsymbol{v})$, where $\texttt{sim}(\cdot, \cdot)$ denotes the cosine similarity. The prediction probability of $\boldsymbol{x}_{test}$ can be denoted by $p(y_i|\boldsymbol{x}_{test}) = \frac{\exp(\texttt{sim}(\boldsymbol{t}_i \cdot \boldsymbol{v}))}{\sum_{i=1}^{K} \exp(\texttt{sim}(\boldsymbol{t}_i \cdot \boldsymbol{v}))}$.

The traditional classifiers learn closed-set visual concepts, whereas the vision-language pre-trained models like CLIP learn open-set visual concepts via the high-capacity text encoder. This leads to a broader semantic space, making learned representation more transferable to downstream tasks. The pre-trained vision-language CLIP model therefore offers impressive zero-shot capabilities. We leverage the frozen CLIP model for image recognition in continual learning settings without any training in a strictly zero-shot fashion. For evaluation, we first initialize both the CLIP encoders and freeze the weight. Different from traditional evaluation where all the test dataset will be fed to the model, in continual evaluation we assume that the model is learning a current task t and evaluated only on the tasks observed so far.

4 EXPERIMENTS

4.1 EXPERIMENTAL PROTOCOLS

Datasets. We evaluate Continual-CLIP model on seven different datasets and 13 different learning task configurations. For class-incremental settings, we evaluate our model on CIFAR-100, ImageNet-100 & 1K, TinyImageNet under different class splits. **(a)** In CIFAR-100, we compare the performance on 10 steps (10 new classes per step), 20 steps (5 new classes per step), and 50 steps (2 new classes per step) (Douillard et al., 2022; Yan et al., 2021). **(b)** In ImageNet-100, we consider two evaluation settings; *ImageNet-100-B0* which has the same number of classes for all the steps (i.e., 10 classes per step) and *ImageNet-100-B50* that contains 50 classes for the first step and the rest of the 50 classes are observed incrementally in the next 10 steps (5 classes per steps) (Yan et al., 2021). **(c)** We divided ImageNet-1K into 10 incremental steps (100 classes per step) (Yan et al., 2021). **(d)** In TinyImageNet, we kept 100 classes for the first (or base) step, and then similar to ImageNet-100, add the rest of the classes incrementally in three different number of steps i.e., 5 steps (20 classes per step), 10 steps (10 classes per step), and 20 steps (5 classes per step) (Zhu et al., 2021). Note that we do not evaluate under task-incremental setting where task IDs are known at inference since it is the easiest of the continual learning settings.

We evaluate our model in domain-incremental setting on CORe50 and CLEAR-10 & 100 datasets. The CORe50 has 11 different scenarios and 10 classes. Out of which we exclude 8 domains since we are not training (or fine-tuning) the model and only consider 3 domains for evaluation following the guidelines in (Lomonaco & Maltoni, 2017). For the experiment on CLEAR-10 & 100, we use the Avalanche (Lomonaco et al., 2021) pre-build methods to get the evaluation scenarios.

We used the Gaussian schedule CIFAR-100 for the evaluation in task-free (or task-agnostic) setting. We use the same setup as (Shanahan et al., 2021) to check the effectiveness of our method in task-agnostic setting. For the evaluation, (Shanahan et al., 2021) keep the evaluation dataset aside and do evaluation on the whole dataset after every task. Since our Continual-CLIP model is frozen, we report the final test accuracy for comparison.

Evaluation Metrics. In class-incremental setting, we compare our methods with other baseline approaches in terms of "Avg" accuracy, which is the average of all accuracy values obtained at each time step and "Last" accuracy which is the final accuracy after learning all the tasks. In domain-incremental setting, we used "In-domain","Next-domain", "Forward transfer", "Backward transfer", and "overall accuracy" to compare our method. Note that for the CORe50 dataset, only a single test set is available, so we report the accuracy on this set. Similarly, for the task-agnostic setting, we only report test accuracy.

Implementation Details. We use the official CLIP (Radford et al., 2021) implementation in zero-shot evaluation settings. To build continual scenarios for class-incremental setting, we heavily used Continuum (Douillard & Lesort, 2021) and follow the same evaluation setting from Douillard et al. (2022). For the domain-incremental scenarios on CORe50 and CLEAR datasets, we use the Avalanche library (Lomonaco et al., 2021). Since the CIFAR100 and TinyImageNet datasets contain low resolution images, we use a "`a bad photo of a {category}`" prompt template. The datasets like ImageNet-100 and 1K have high resolution images so we use "`a good photo of a {category}`" for our evaluation of CLIP for all our experiments.

For the evaluation, we used the same data-reprocessing technique as defined in (Radford et al., 2021), which include Bicubic interpolation, Center Cropping, and Normalization.

4.2 RESULTS

Class-incremental Setting. The results in Table 1 compare our method with other baselines on CIFAR-100 dataset in three different settings (10, 20 and 50 steps). In 10 steps setting, even without any training or fine-tuning, Continual-CLIP achieves competitive results in terms average and last accuracy, compared with the recent state-of-the-art methods such as DyTox (Douillard et al., 2022) and DER (Yan et al., 2021). Specifically, In 20 steps setting, Continual-CLIP reaches 75.95% in "Avg" accuracy, and for the 50 steps setting, it reaches 76.49% in "Avg" accuracy. The last accuracy is same for all the cases (since it is zero-shot evaluation). The results suggest that Continual-CLIP

Table 1: Comparison of state-of-the-art CL methods on CIFAR100 benchmark in class-incremental setting, in terms of the average and last task accuracy values.

	10 steps		20 steps		50 steps	
Methods	**Avg**	**Last**	**Avg**	**Last**	**Avg**	**Last**
iCaRL (Rebuffi et al., 2017)	65.27	50.74	61.20	43.74	56.08	36.62
UCIR (Hou et al., 2019)	58.66	43.39	58.17	40.63	56.86	37.09
BiC (Wu et al., 2019)	68.80	53.54	66.48	47.02	62.09	41.04
RPSNet (Rajasegaran et al., 2019b)	68.60	57.05	-	-	-	-
WA (Zhao et al., 2020)	69.46	53.78	67.33	47.31	64.32	42.14
PODNet (Douillard et al., 2020)	58.03	41.05	53.97	35.02	51.19	32.99
DER (w/o P) (Yan et al., 2021)	**75.36**	65.22	74.09	62.48	72.41	59.08
DER (Yan et al., 2021)	74.64	64.35	73.98	62.55	72.05	59.76
DyTox (Douillard et al., 2022)	67.33	51.68	67.30	48.45	64.39	43.47
DyTox+ (Douillard et al., 2022)	74.10	62.34	71.62	57.43	68.90	51.09
Continual-CLIP	75.17	**66.72**	**75.95**	**66.72**	**76.49**	**66.72**

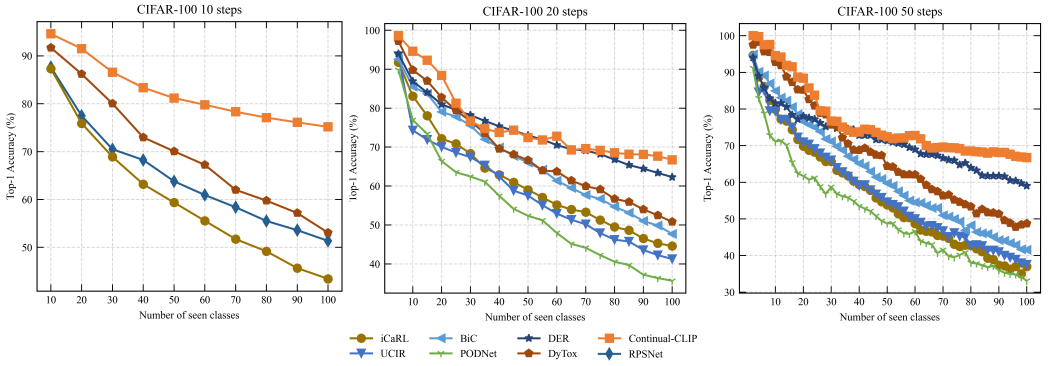

Figure 2: Step-wise performance trends on the CIFAR-100 for 10, 20, and 50 steps. Continual-CLIP performs well especially for longer incremental episodes. Note that the close competitor DER expands the architecture with new tasks, thereby significantly increasing compute complexity.

consistently performs better than majority of the compared methods by a significant margin. This trend is especially notable on large number of tasks, which is a harder case in CL (Figure 2).

The results on the ImageNet-100 & 1K datasets are presented in Table 2. In standard settings (i.e., same number of classes in all steps; see columns 2-5), Continual-CLIP shows an improvement of 7.84% on ImageNet-100 and 4.63% on ImageNet-1K dataset in terms of the "Avg" accuracy compared with the second best method. Similarly, a significant improvement of 7.71% on ImageNet-100 and 7.70% on ImageNet-1K in "Last" accuracy is achieved. For the other setting with 50 base classes (i.e., ImageNet-100-B50), Continual-CLIP shows better results than the recent methods.

From the results reported in Table 3, we observe that on TinyImageNet dataset with base class of 100, Continual-CLIP shows consistent improvements in all 3 settings. Continual-CLIP shows on average 20.30% improvement compared with the second best method DyTox in terms of the average accuracy, also it shows impressive gains in terms of the last accuracy.

Domain-incremental setting. Table 4 shows the comparison of our method with the winning entry of the CVPR 2022 CLEAR Challenge Leaderboard (AIcrowd, 2022; Lin et al., 2021). The values are reported on the test set following the evaluation protocol in (Lomonaco & Maltoni, 2017). For CLEAR-10 & 100, Continual-CLIP performs competitively with the top performing team.

Task-agnostic setting. This is a general setting in which there is no constraint of task, class, or domain and is considered to be quite challenging. This setting is relatively under-explored in the literature. In Table 6, we show a comparison of our method with the previous best method i.e., Encoders and Ensemble (Shanahan et al., 2021). Continual-CLIP outperforms the compared approach with a large margin without any need of training or model ensembles.

Table 2: Comparison of state-of-the-art CL methods on different ImageNet benchmarks, in class-incremental settings with 10 splits, in terms of average and last accuracy values.

Methods	ImageNet100-B0		ImageNet1K		ImageNet100-B50	
	Avg	**Last**	**Avg**	**Last**	**Avg**	**Last**
iCaRL (Rebuffi et al., 2017)	-	-	38.40	22.70	-	-
UCIR (Hou et al., 2019)	-	-	-	-	68.09	57.30
WA (Zhao et al., 2020)	-	-	65.67	55.60	-	-
TPCIL (Tao et al., 2020)	-	-	-	-	74.81	66.91
PODNet (Douillard et al., 2020)	-	-	-	-	74.33	-
Simple-DER (Li et al., 2021b)	-	-	66.63	59.24	-	-
DER (w/o P) (Yan et al., 2021)	77.18	66.70	68.84	60.16	78.20	74.92
DER (Yan et al., 2021)	76.12	66.06	66.73	58.62	77.13	72.06
DyTox (Douillard et al., 2022)	73.96	62.20	-	-	-	-
DyTox+ (Douillard et al., 2022)	77.15	67.70	70.88	60.00	-	-
Continual-CLIP	**85.00**	**75.42**	**75.51**	**67.71**	**79.69**	**75.42**

Table 3: Comparison of state-of-the-art CL methods on different TinyImageNet splits in class-incremental settings with 50 base classes, in terms of the average and last accuracy values.

Methods	5 steps		10 steps		20 steps	
	Avg	**Last**	**Avg**	**Last**	**Avg**	**Last**
EWC (Kirkpatrick et al., 2017)	19.01	6.00	15.82	3.79	12.35	4.73
LwF (Li & Hoiem, 2017)	22.31	7.34	17.34	4.73	12.48	4.26
LwF-MC (Rebuffi et al., 2017)	29.09	15.63	23.03	13.25	17.31	7.95
iCaRL (Rebuffi et al., 2017)	34.27	23.22	30.94	20.82	27.83	20.16
iCaRL-NCM (Rebuffi et al., 2017)	45.95	34.60	43.22	33.22	37.85	27.54
EEIL (Castro et al., 2018)	47.17	35.12	45.03	34.64	40.41	29.72
UCIR (Hou et al., 2019)	50.30	39.42	48.58	37.29	42.84	30.85
MUC (Liu et al., 2020)	32.23	19.20	26.67	15.33	21.89	10.32
PASS (Zhu et al., 2021)	49.54	41.64	47.19	39.27	42.01	32.93
DyTox (Douillard et al., 2022)	55.58	47.23	52.26	42.79	46.18	36.21
Continual-CLIP	**70.49**	**66.43**	**70.55**	**66.43**	**70.51**	**66.43**

In summary, our extensive empirical evaluations and comparisons provide an evidence that the Continual-CLIP consistently shows impressive results in all continual learning settings, without the need of any fine-tuning (or training), dedicated memory for past exemplars and model copies, complex hyper-parameter tuning, dynamic model expansion or changing classification heads.

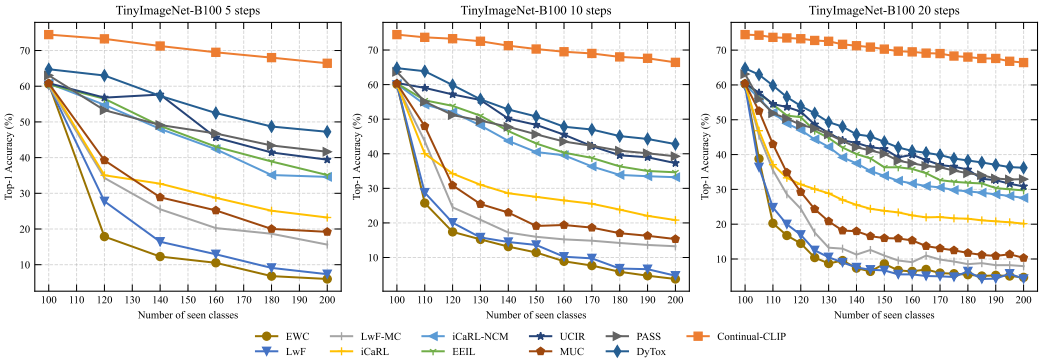

Figure 3: Accuracy trends on the TinyImageNet dataset in three different settings of step sizes. Note that the other competing methods require training at each stage, use memory buffers, may not apply to all CL settings and/or dynamically expand the architecture to learn new tasks.

Table 4: Domain incremental setting comparison with winning team of the recent CVPR 2022 Continual LEArning on Real-World Imagery (CLEAR) Challenge (AIcrowd, 2022; Lin et al., 2021).

Datasets	Methods	Overall Acc	Next-domain	In-domain	Backward	Forward
CLEAR-10	Top-1 team	92.70	92.50	93.40	**94.20**	90.90
	Continual-CLIP	**93.79**	**93.51**	**93.58**	93.83	**93.33**
CLEAR-100	Top-1 team	91.46	91.25	91.99	93.40	89.20
	Continual-CLIP	**93.63**	**93.50**	**93.50**	**93.66**	**93.33**

Table 5: Core50 dataset comparisons with other baselines in domain incremental setting. The values for compared methods are from (Lomonaco & Maltoni, 2017; Parisi et al., 2018).

Methods	Test Acc (%)
Naïve (Lomonaco & Maltoni, 2017)	54.69
EWC (Kirkpatrick et al., 2017)	57.40
LwF (Zhao et al., 2020)	59.42
Cumulative (Lomonaco & Maltoni, 2017)	65.15
GDM (Parisi et al., 2018)	74.87
Continual-CLIP	**84.73**

4.3 Analysis on Text Prompts

In this section, we investigate the effect of different class names used as part of text prompts on the Continual-CLIP accuracy's. We found different types of naming convention for the ImageNet-1K dataset classes, and use them to evaluate our model's performance. Specifically, we used three names from three different sources: (**a**) ImageNet original labels, where each of the 1000 ImageNet (Deng et al., 2009) classes correspond to a WordNet synset (a set of synonyms), (**b**) the original ImageNet labels have overlapping meaning (e.g., "`nail`", CLIP model understood it has "`fingernail`" so it was changed to "`metal nail`") so Radford et al. (2021) curated the default labels to overcome this confusion, and (**c**) the labels with the first synonym from each Synset. Table 7 shows the effect of different class names conventions on model's performance. With class name type (**b**) model achieves better results than the other two class types, this is because the class names for (**b**) is curated in such a way that there will be a clear boundary and distinction between different class names, and for other two types there can be overlapping class meanings.

We further explore the effectiveness of our method with the prompt engineering. For that, we used different prompts and analyse the results on ImageNet-100 dataset. In Table 8, we report the results for two different prompt techniques for continual zero-shot CLIP. **(p1) Decision-based pooling** - We compute the score for each prompt separately and then average pool them. **(p2) Embedding pooling** - We first compute the text-embedding for each prompt by passing it through the text-encoder and then create a single classifier head by stacking all the embeddings. The results in Table 8 show the trade-off between embedding and decision-based pooling. Embedding and Decision-based pooling needs to collect several prompt templates and requires additional computation to get the text-embedding for each prompt as compared to our main results. Also, with the decision and embedding pooling, we need domain expert knowledge for prompt-engineering, and repeat the process over multiple trials until we can find best performing prompt for the given scenario.

Table 6: Comparisons in domain incremental setting on Core50 dataset. Results for compared methods are from (Lomonaco & Maltoni, 2017; Parisi et al., 2018)

Methods	Test Acc (%)
Encoders and Ensemble (Shanahan et al., 2021)	39.0
Continual-CLIP	**66.72**

Table 7: Effect of using different class names on the Continual-CLIP accuracy on ImageNet-1k with a prompt "`a photo of a {}`."

Class Names Type	Avg Acc (%)	Last Acc (%)
(a) ImageNet default	72.96	64.44
(b) Radford et al. (2021) curated	74.81	66.58
(c) First synomym from each subset	71.65	63.97

Table 8: The performance of Continual-CLIP on 80 different prompts with `(p1) decision-based pooling` (the average of all 80 prompts) and `(p2) embedding pooling` on ImageNet-100 dataset.

Class Names Type	Acc (%)	Last Acc (%)
Decision-based pooling	82.24	72.11
Decision-based pooling (top-10)	84.63	74.94
Embedding pooling	84.85	75.46

5 Conclusion

In this work, we introduce a simple yet effective baseline that consistently achieves favorable results for three different incremental learning settings; (1) class-incremental, (2) domain-incremental, and (3) task-agnostic incremental learning. Our Continual-CLIP is a standard CLIP model evaluated to work in continual learning scenarios. The experiments on different benchmarks with challenging configurations shows that the Continual-CLIP outperforms the current state-of-the-art methods in continual learning without the need of any fine-tuning, replay buffers, or memory overhead. Continual-CLIP model can be used in any continual learning setting with zero or little modification. The state-of-the-art performance on the large datasets like ImageNet-1k (in class-incremental), CORe50 (in domain-incremental) shows that Continual-CLIP is scalable.

In future work, our simple baseline approach built on top of zero-shot transfer capabilities of CLIP can be extended with fast adaptation methodologies for downstream continual tasks. Current work provides grounds for the future development in the continual learning leveraging the vision-language foundational models. Although such models show excellent performance, there arise new questions based on their behaviour, *e.g.*, it can be seen from the confusion matrix in Figure 4 (left: CIFAR100) the model gets confused when there is a close semantic resemblance: the class name for the class index "`50`" is "`mouse`" and the model is predicts it as "`74`" which is "`shrew`". Our work also motivates rethinking the progress made so far in the continual learning problems where the state of the art methods come with several constraints and promotes looking for generic solutions that transcend beyond narrow settings and cumbersome memory and compute requirements.

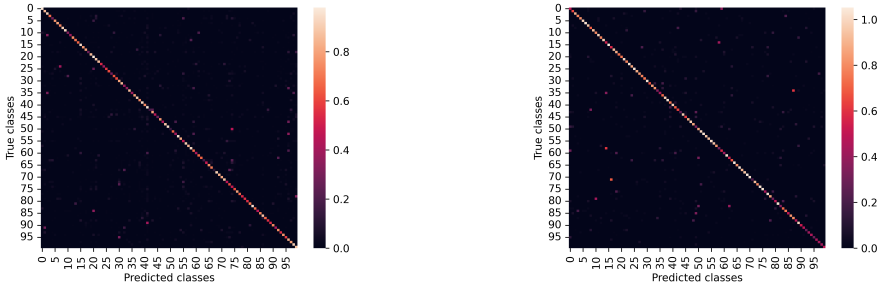

Figure 4: *Left:* Confusion matrix for CIFAR-100 dataset. *Right:* Confusion matrix for ImageNet-100 dataset. We note that majority errors occur between semantically similar classes.

REFERENCES

AIcrowd. Cvpr 2022 clear challenge: Leaderboards, 2022. URL https://www.aicrowd.com/challenges/cvpr-2022-clear-challenge/leaderboards.

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

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

Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *In Proceedings of the European conference on computer vision*, pp. 233–248, 2018.

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

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Arthur Douillard and Timothée Lesort. Continuum: Simple management of complex continual learning scenarios, 2021.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *In Proceedings of the European Conference on Computer Vision*, pp. 86–102. Springer, 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*, pp. 9285–9295, June 2022.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 831–839, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Nitin Kamra, Umang Gupta, and Yan Liu. Deep generative dual memory network for continual learning. *arXiv preprint arXiv:1710.10368*, 2017.

Zixuan Ke, Bing Liu, and Xingchang Huang. Continual learning of a mixed sequence of similar and dissimilar tasks. *Advances in Neural Information Processing Systems*, 33:18493–18504, 2020.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021.

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.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021a.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019a.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

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

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

Zhuoyun Li, Changhong Zhong, Sijia Liu, Ruixuan Wang, and Wei-Shi Zheng. Preserving earlier knowledge in continual learning with the help of all previous feature extractors, 2021b. URL `https://arxiv.org/abs/2104.13614`.

Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. The clear benchmark: Continual learning on real-world imagery. In *Conference on Neural Information Processing Systems*, 2021.

Yu Liu, Sarah Parisot, Gregory Slabaugh, Xu Jia, Ales Leonardis, and Tinne Tuytelaars. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. In *In Proceedings of the European Conference on Computer Vision*, pp. 699–716. Springer, 2020.

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

Vincenzo Lomonaco, Lorenzo Pellegrini, Andrea Cossu, Antonio Carta, Gabriele Graffieti, Tyler L. Hayes, Matthias De Lange, Marc Masana, Jary Pomponi, Gido van de Ven, Martin Mundt, Qi She, Keiland Cooper, Jeremy Forest, Eden Belouadah, Simone Calderara, German I. Parisi, Fabio Cuzzolin, Andreas Tolias, Simone Scardapane, Luca Antiga, Subutai Amhad, Adrian Popescu, Christopher Kanan, Joost van de Weijer, Tinne Tuytelaars, Davide Bacciu, and Davide Maltoni. Avalanche: an end-to-end library for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2nd Continual Learning in Computer Vision Workshop, 2021.

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

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

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

German I Parisi, Jun Tani, Cornelius Weber, and Stefan Wermter. Lifelong learning of spatiotemporal representations with dual-memory recurrent self-organization. *Frontiers in neurorobotics*, 12: 78, 2018.

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

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

Jathushan Rajasegaran, Munawar Hayat, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Ming-Hsuan Yang. An adaptive random path selection approach for incremental learning. *arXiv preprint arXiv:1906.01120*, 2019b.

Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *Conference on Neural Information Processing Systems*, 2022.

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

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

Murray Shanahan, Christos Kaplanis, and Jovana Mitrović. Encoders and ensembles for task-free continual learning. *arXiv preprint arXiv:2105.13327*, 2021.

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

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

CS231N Course Stanford. Tiny imagenet challenge. URL `http://cs231n.stanford.edu/tiny-imagenet-200.zip`.

Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. Topology-preserving class-incremental learning. In *In Proceedings of the European Conference on Computer Vision*, pp. 254–270. Springer, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Zifeng Wang, Tong Jian, Kaushik Chowdhury, Yanzhi Wang, Jennifer Dy, and Stratis Ioannidis. Learn-prune-share for lifelong learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 641–650. IEEE, 2020.

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 Proceedings of the European Conference on Computer Vision*, 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*, pp. 139–149, 2022b.

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

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

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

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

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

Tingting Zhao, Zifeng Wang, Aria Masoomi, and Jennifer Dy. Deep bayesian unsupervised lifelong learning. *Neural Networks*, 149:95–106, 2022.

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