vCLIMB: A Novel Video Class Incremental Learning Benchmark

Andrés Villa[1,2*], Kumail Alhamoud[2], Victor Escorcia[3], Fabian Caba Heilbron[4],
Juan León Alcázar[2], Bernard Ghanem[2]

[1]Pontificia Universidad Católica de Chile, [2]King Abdullah University of Science and Technology (KAUST),
[3]Samsung AI Center Cambridge, [4]Adobe Research

afvilla@uc.cl, kumail.hamoud@kaust.edu.sa, v.castillo@samsung.com
caba@adobe.com, {juancarlo.alcazar, bernard.ghanem}@kaust.edu.sa

Abstract

Continual learning (CL) is under-explored in the video domain. The few existing works contain splits with imbalanced class distributions over the tasks, or study the problem in unsuitable datasets. We introduce vCLIMB, a novel video continual learning benchmark. vCLIMB is a standardized test-bed to analyze catastrophic forgetting of deep models in video continual learning. In contrast to previous work, we focus on class incremental continual learning with models trained on a sequence of disjoint tasks, and distribute the number of classes uniformly across the tasks. We perform in-depth evaluations of existing CL methods in vCLIMB, and observe two unique challenges in video data. The selection of instances to store in episodic memory is performed at the frame level. Second, untrimmed training data influences the effectiveness of frame sampling strategies. We address these two challenges by proposing a temporal consistency regularization that can be applied on top of memory-based continual learning methods. Our approach significantly improves the baseline, by up to 24% on the untrimmed continual learning task. The code of our benchmark can be found at: https://vclimb.netlify.app/.

1. Introduction

Deep neural networks rely on large-scale datasets to achieve state-of-the-art performance on modern computer vision tasks [6, 10, 18, 20]. A significant result of such pretraining is enabling feature reuse [12, 41], by means of fine-tuning the learned weights for smaller downstream tasks [3, 22, 28]. Due to legal or technical constraints, and the fact that labeling data is expensive and time-consuming [34], real-world deep-learning pipelines would rarely involve a single fine-tuning stage. Instead, these pipelines could require the sequential fine-tuning of large models in a set of independent tasks that are learned sequentially. Under these conditions, deep neural networks suffer from what is known as *catastrophic forgetting* [13], where the fine-tuning on novel tasks significantly reduces the performance of the model in a previously learned task, and *drift* [29], where unseen training data does not fit the previously estimated class distribution. Continual learning [11] directly models such a scenario, by adapting a neural network model into a sequential series of tasks. We focus on a special case of CL: *class incremental learning (CIL)*, where the labels and data are mutually exclusive between tasks, training data is available only for the current task, and there are no tasks ids.

With 500 hours of video from diverse categories uploaded to YouTube every minute and a billion people actively using TikTok every month [4, 36], video content of varying quality is available at an unprecedented scale. With such large volumes of data, it is important to develop models that can effectively learn from continuous streams of untrimmed video data. Remarkably, few research efforts have addressed continual learning with video [24, 26, 44]. Despite these current works, video continual learning methods still show a large variability in their experimental protocols, making direct comparisons hard to establish. These protocols present the following limitations. **(1)** They are not publicly available. **(2)** They do not explore a realistic setup with untrimmed videos. **(3)** Most of them leverage a large pretraining step, which warms up the model by learning a sample of classes from the same distribution and is not always available in a real continual learning scenario.

We directly address these limitations and propose vCLIMB (**v**ideo **CL**ass **I**ncre**M**ental Learning **B**enchmark), a novel benchmark devised for the evaluation of continual learning in video. Our test-bed defines a fixed task split on the original training and validation sets of three well known video datasets: UCF101 [33], ActivityNet [5] and Kinetics [6]. vCLIMB follows the standard class incremental continual learning scenario, but includes some modifications to better fit the nature of video data in human ac-

*Work done during an internship at KAUST.

tion recognition tasks. First, to achieve fair comparisons between video CIL methods that use memory, we re-define *memory size* [27, 38] to be the total number of frames, instead of the total number of video instances, that the memory can store. We report this as *Memory Frame Capacity* in our tables to avoid confusion with the memory size defined in image CIL. This means we are not only concerned about selecting the best videos to store in memory, but we also want to identify the best set of frames to keep in memory. Second, since fine-grained temporal video annotations are expensive (especially for long videos), we analyze the effect of using trimmed and untrimmed video data in continual learning. To the best of our knowledge, this is the first work to explore continual learning with untrimmed videos.

Using vCLIMB's data splits, we establish an initial set of baselines by adapting well-known continual learning methods [1, 17, 27, 38] from the image recognition domain into the activity recognition domain. After benchmarking these baseline methods, we propose a novel strategy for video continual learning that leverages the inherent temporal consistency in video data to better approach the continual action recognition problem. We believe that vCLIMB's standardized approach to prototyping and evaluating video continual learning will enable future works in this area.

Contributions. This paper proposes vCLIMB, a novel benchmark for continual learning in video, which focuses on the activity classification task. Our work brings the following contributions: **(1)** A standardized benchmark for continual learning in video action recognition, which defines the training protocols and associated metrics for three video datasets; this novel continual learning setup includes a more realistic combination of trimmed and untrimmed videos. **(2)** We re-purpose and evaluate four baseline methods from the image domain into the video domain. **(3)** We present a novel strategy based on consistency regularization that can be built on top of memory-based methods to reduce memory consumption while improving performance.

2. Related Work

Continual Learning (CL) studies methods that can learn novel concepts continually without forgetting previous knowledge. In the class incremental learning (CIL) setting, models are trained sequentially on a set of tasks. Each incremental task consists of labeled data from novel classes, which are learned individually without considering the data from previous tasks. CIL approaches can be grouped into two broad categories: regularization-based and memory-based methods. While regularization-based methods penalize abrupt changes in the most relevant parameters learned from previous tasks, memory-based methods mitigate catastrophic forgetting by retaining a limited amount of training instances from previous tasks. Although there is extensive literature on continual learning [1, 7, 8, 15, 17, 21, 25, 27, 31, 42], we restrict our review to a sample of the most widely adopted baselines.

Regularization-Based Methods. Regularization techniques try to keep constant the weights that are important for the previous tasks. They differ in how they estimate the importance value of the model parameters learned in previous tasks. These values are commonly updated and stored in an importance matrix at the end of each task. Elastic Weight Consolidation (EWC) [17] uses the Fisher Information Matrix to make that estimation. Memory Aware Synapses (MAS) [1] estimates the importance of each parameter in a self-supervised manner by measuring how small changes in the parameters affect the output of the model. MAS is of particular interest in the video domain because it can handle weakly labeled instances. These methods include a regularization factor λ_{reg}, which controls how relevant the previously learned tasks are. For a large factor, the model will prioritize the previous tasks over the current one.

Memory-Based Methods. Memory-based methods select and store samples into a memory buffer for future replay. This memory has a limited size and is available when learning a new task. While a naive baseline randomly chooses instances from previously learned classes, current strategies attempt to select the best subset of training samples per class to move into the buffer. We focus on two representative methods that have been shown to work well with high-resolution image datasets. Incremental Classifier and Representation Learning (iCaRL) [27] combines rehearsal and distillation strategies. During training, it selects the most representative instances per class following a nearest-mean-of-exemplars rule. The Bias Correction method (BiC) [38] follows iCaRL's instance sampling approach, but augments the classification layer with a bias correction layer. Bias correction mitigates the imbalance between the large amount of data from the new task and the relatively scarce data from previous tasks, which is only available in memory.

Consistency Regularization. Consistency regularization techniques are used to ensure that a model's output is invariant to various augmentations, which is common in semi-supervised learning [2, 19, 30, 32, 39]. For example, such methods have been shown to improve the generation of images from a few samples [43, 45]. Our work investigates catastrophic forgetting in video continual learning and proposes a consistency loss to help memory-replay methods significantly alleviate this impeding nuisance.

Continual Learning in Video. Despite the growing interest in CL in the image domain, the first three works to report results on video data were only recently published. Zhao *et al*. [44] proposed a spatio-temporal knowledge transfer strategy to mitigate catastrophic forgetting. A concurrent work [26] estimated the subset of feature channels that contributes the most to the predictions of the previous tasks,

Set	In-distribution Pretraining	Tasks	Videos Per Task			Classes Per Task	Avg. Frames Per Video	Untrimmed Video
			Train	Val	Test			
i-Something-Something-B0 [44]	None	4	–	–	–	10	–	✗
i-Something-Something-B20 [44]	20 classes	5	–	–	–	4	–	✗
i-Kinetics-B0 [44]	None	4	–	–	–	10	–	✗
i-Kinetics-B20 [44]	20 classes	5	–	–	–	4	–	✗
UCF101-50 [26]	51 classes	5/10/25	–	–	–	10/5/2	–	✗
HMDB51-25 [26]	26 classes	5/25	–	–	–	5/1	–	✗
Something-Something V2-90 [26]	84 classes	9/18	–	–	–	10/5	–	✗
vCLIMB UCF101	None	10	928	131	272	10	183	✗
vCLIMB UCF101	None	20	464	65	136	5	183	✗
vCLIMB Kinetics	None	10	24628	1988	3977	40	250	✗
vCLIMB Kinetics	None	20	12314	994	1988	20	250	✗
vCLIMB ActivityNet-Untrim.	None	10	1001	492	–	20	3542	✓
vCLIMB ActivityNet-Untrim.	None	20	500	246	–	10	3542	✓
vCLIMB ActivityNet-Trim.	None	10	1541	765	–	20	3879	✗
vCLIMB ActivityNet-Trim.	None	20	770	383	–	10	3879	✗

Table 1. **CIL Benchmark Statistics.** In vCLIMB we provide 8 splits for class incremental learning, each split contains 10 or 20 tasks. Our Kinetics and ActivityNet setups contain large-scale video data and provide long sequences of tasks with no video pretraining. This makes our splits more suitable for measuring forgetting. We highlight that ActivityNet-Untrim provides a realistic challenge to test a model's ability to continually learn from weakly labeled video streams.

and introduced a temporal mask that keeps this subset stable while learning a new task. It also included a distillation loss, which allows only the least relevant feature maps to be updated while learning a new task. Finally, Ma *et al*. [24] also approached the class incremental video classification problem by regularizing the feature space in consecutive tasks.

Although existing works bootstrapped the study of continual learning in video data, all of them use different evaluation protocols, making direct comparisons between methods difficult. Moreover, these works propose to pretrain the model with a large set of classes of the same data distribution (up to half of the total), as shown in Table 1. Such an arrangement is unnatural for continual learning, as it makes it difficult to disentangle the effects of catastrophic forgetting. In contrast, our benchmark presents a more challenging and realistic setup. Our splits contain up to 20 tasks, with a balanced number of classes per task. Furthermore, vCLIMB includes three video datasets, which enables the study of the continual learning problem in more diverse scenarios. Finally, previous works do not provide a detailed analysis of the proposed video continual learning setups. In our benchmark, we provide extensive empirical evaluations to analyze individual splits and identify unique properties of continual learning in video, including the memory size and the adoption of untrimmed video configurations.

3. vCLIMB: A Video Class Incremental Learning Benchmark

Notation & Problem Definition Similar to the image domain, we train a single neural network (F_ω) with parameter set (ω) over a sequence of tasks (R_n). Each task in the sequence contains its own training data $R_i = \{(X_0, Y_0), (X_1, Y_1)...(X_n, Y_n)\}$, with X_n the input data and Y_n its corresponding ground truth. We optimize the parameter set ω in F sequentially over R_n, searching for $\bar{F}_\omega$ that maximizes the average accuracy over R_n.

Datasets and Tasks. The bottom half of Table 1 summarizes the main attributes of vCLIMB. We design vCLIMB on top of three well-known datasets for action recognition: **(1)** UCF101 [33], which has 13.3K videos from 101 classes, **(2)** Kinetics [6], a large-scale video dataset with over 300K short clips distributed over 400 action classes, and **(3)** ActivityNet [5], which can be used for trimmed and untrimmed activity classification and contains 20K videos from 200 activity classes. We utilize the diversity of videos in ActivityNet and provide two subsets: ActivityNet-Trim, in which every part of the videos when an action happens is considered as an independent video, and ActivityNet-Untrim, which is more challenging and labels a whole video with its most representative action class. We create two different CIL sequences of tasks for each dataset. The first sequence contains ten tasks, and the second sequence contains twenty tasks. Statistics about the number of classes per task, the number of videos per task, and the average number of frames in each split are provided in Table 1.

Metrics for Video Continual Learning (CL). In vCLIMB, we use the standard CL metrics: Final Average Accuracy (Acc) and Backward Forgetting (BWF). Acc is the average classification accuracy of the model evaluated on all learned tasks, including the last task it was trained on [23, 37]. This metric is essential to show how the average performance of the model degrades as it learns new tasks. BWF complements Acc and measures the influence of the learned task i

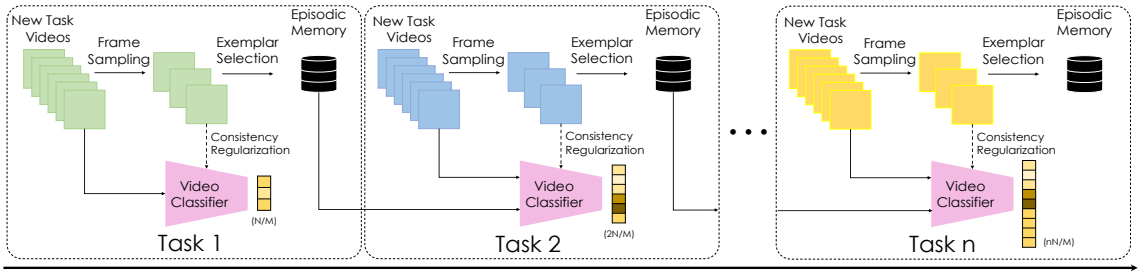

Figure 1. **Outline of our Memory-based CIL setting.** In video CIL, a model sequentially learns a set of video classification tasks. Memory-based methods define a limited episodic memory to store a few temporally down-sampled examples of previous tasks. We propose a consistency regularization that helps the model to better remember previous tasks from down-sampled examples in memory.

in the performance of the previous tasks [23], as:

$$BWF_i = \frac{1}{N_i - 1} \sum_{j=1}^{N_i - 1} (R_{j,j} - R_{N_i,j}) \qquad (1)$$

where N_i is the number of learned tasks after learning the task i, and $R_{j,j}$ and $R_{N_i,j}$ represents the accuracy on the task j after learning the task j, and learning a new task the task i, respectively. Given a total of N tasks, we report the final Backward Forgetting $BWF = BWF_N$ in our tables.

3.1. Unique Challenges of Video CL

Video CIL comes with unique challenges. (1) Memory-based methods developed in the image domain are not scalable to store full-resolution videos, so novel methods are needed to select representative frames to store in memory. (2) Untrimmed videos have background frames that contain less helpful information, thus making the selection process more challenging. (3) The temporal information is unique to video data, and both memory-based and regularization-based methods need to mitigate forgetting while also integrating key information from this temporal dimension. These challenges informed the design choices of vCLIMB.

Re-defining Memory Size. Unlike image benchmarks for class incremental learning (CIL), our video instances contain a temporal dimension whose size could show large variability. To favor fair comparisons between methods, we define the working memory size in terms of stored frames. This design choice avoids a scenario where longer videos are preferred (or even trivially selected) to maximize the amount of data stored in the working memory. Moreover, it creates a new unique aspect of CIL in video data, as methods must decide first what subset of frames should be selected, and then decide what video to store according to the sub-sampled videos. We follow the same instance per class ratio as [27], which is 20 when the model has learned all training classes. Therefore, we define a memory that can save at most 8000, 4000, and 2020 videos for Kinetics, ActivityNet, and UCF101, respectively. If we store videos in memory without down-sampling them, this would correspond to saving 3.25%, 25.95%, and 21.76% of the total frames in Kinetics, ActivityNet, and UCF101. As we show later, these storage requirements can be significantly reduced using temporal consistency regularization.

Untrimmed Video Data for CIL. In the untrimmed classification setting, the action of interest may occur at any time in the video and its boundaries are not known. This problem formulation has no direct counterpart in the image classification domain. The annotation scheme of ActivityNet allows us to analyze this scenario for class incremental learning. In ActivityNet, videos contain one or multiple temporal segments defining the occurrence of an action instance, while unlabeled segments constitute a background set where no relevant action takes place [5].

We leverage this unique property of video data and define two independent setups for video CIL. In the trimmed setup, we only use frames that belong to a labeled action segment. In the untrimmed setup, we freely sample frames from the whole video, regardless of whether they belong to the main action or the background. For consistency in the untrimmed scenario, we give every frame in the video the same label. We select a primary label as the action with the longest temporal support in the video, and discard any video that contains instances of 2 or more different labels. We empirically find that this assignment only discards 0.15% of the ActivityNet dataset. This untrimmed learning task more closely resembles the real world scenario of CIL, where a model learns from a continuous stream of diverse videos. Given the scale of current video services and the costly nature of fine-grained labels, real models would likely be learning from a stream with weak video-level annotations.

Baselines We implement and evaluate these four continual learning methods to serve as baselines [1, 17, 27, 38] because they are widely used and easily scalable to the video domain. We also compare with a naive memory-based strategy, which selects samples randomly for memory creation. We provide the implementations for these methods as part of our video CIL benchmark.

3.2. A Stronger Baseline with Temporal-consistency Regularization

We present a novel strategy for CIL in the video domain. Our approach relies on one of the unique characteristics of

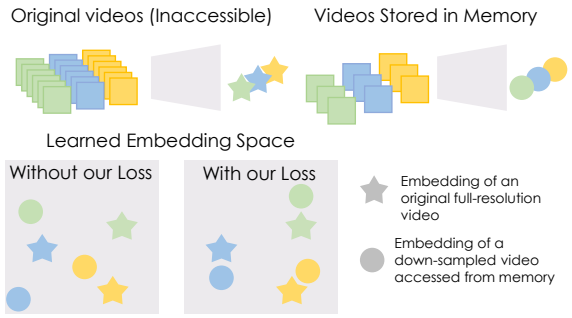

Figure 2. **Temporal Consistency Loss.** Our consistency loss encourages learning representations that are robust to frame sampling, enabling the CIL model to remember old tasks by looking only at down-sampled versions of those videos.

video data, temporal resolution consistency. Figure 1 illustrates our pipeline for memory-based video CIL methods. After completing the training on videos from the first task, the model has the option to select a few temporally subsampled examples to store in the episodic memory. When learning the second task, the model is trained on both the new task videos and a few past-task examples it retained in memory. We introduce a regularization loss, represented by the dashed arrow in Figure 1, during the fine-tuning phase of every new task. This loss constrains the network to estimate similar feature representations for the original video clip and a temporally down-sampled version of it.

This constraint aims to optimize the effectiveness of the data stored in the episodic memory, by enforcing a similar representation between the original clip and its temporally down-sampled version. As shown in Figure 2, the most relevant aspects of our loss is that it reduces the drift between the samples stored in memory and the original samples used at training time. While simple, our constraint directly addresses one of the key aspects of CIL in video: its episodic memory is composed of frame sets instead of full-resolution video clips. In practice, we enforce similar representations by adding a regularization term to the loss function.

Temporal Consistency Loss. When training on a new task, each video (X) will have an augmented version (X^d) by means of temporal down-sampling. We use pairs of X and X^d to calculate a consistency loss L_c over a single forward pass of the network F:

$$L_c = (1-\lambda)L_{cls}(F(X), Y) + \lambda L_{cls}(F(X^d), Y), \quad (2)$$

where L_{cls} is the cross-entropy loss, Y is the ground truth label of X, and λ is the consistency regularization factor.

Our consistency regularization strategy is model agnostic and thus can be adapted to any backbone and CIL memory-replay strategy. Since we use the same set of weights (F) for both loss terms, our method only introduces a linear time increase in the total number of FLOPs.

4. Experimental Evaluation

We now proceed with the experimental assessment of the class incremental continual learning task in vCLIMB. In this section, we first discuss the details of the re-implementation of image methods for continual learning in the video domain. Then, we proceed with the empirical assessment of these baseline methods on the three datasets included in the vCLIMB benchmark: UFC101 [33], Kinetics [6] and ActivityNet [5]. For the ActivityNet dataset, we also evaluate the CIL task in the presence of trimmed and untrimmed video annotations. Moreover, we assess the effectiveness of the proposed regularization method in all the previous scenarios and datasets.

Implementation Details. The memory-based [27,38] and regularization-based [1, 17] baselines are trained for 50 and 20 epochs, respectively, following a fully-supervised setup on each CIL task. We use TSN [35] with a ResNet-34 backbone pretrained on ImageNet. We follow the same temporal data augmentation proposed in [35], using $N = 8$ segments per video. We optimize our model using Adam [16] with a learning rate of 1×10^{-3}. For the temporal consistency loss factor, we use $\lambda = 0.5$. We run MAS [1] with a regularization factor λ_{reg} of 3×10^5. We found EWC [17] to be more challenging to tune for video CIL. The best results are obtained by picking a different regularization factor λ_{reg} for EWC on each individual dataset: 3×10^3 for UCF101, 5×10^2 for Kinetics, and 3×10^5 for ActivityNet.

4.1. Baselines for Video CIL

As outlined in 3.1 the datasets in vCLIMB differ in scale, so we set a different working memory limit for each dataset according to their total number of frames. For each dataset, we perform experiments on two different splits: a 10-task split and a 20-task split. In Table 2, we report the results obtained using the best hyper-parameters for each method.

Consistent with image class incremental learning [14], the regularization-based methods EWC [17] and MAS [1] lag significantly behind replay-based methods, regardless of the difficulty of the dataset or the number of tasks. This is because regularization methods only penalize changes to the model parameters, and thus experience an unavoidable trade-off between learning new tasks and forgetting older tasks. If a larger regularization parameter is used to emphasize learning new tasks, forgetting increases and the average accuracy on old tasks is compromised. If a smaller regularization parameter is used to emphasize remembering old tasks, the accuracy on newer tasks is impaired. Given the difficulty of video CIL, this limitation highlights the importance of future exploration of more sophisticated strategies of memory-free approaches.

Surprisingly, no memory-based approach is superior across all video datasets. While the naive baseline, which

Model	Num. Task	Kinetics				ActivityNet-Trim				UCF101			
		Mem. Video Instances	Mem. Frame Capacity	Acc ↑	BWF ↓	Mem. Video Instances	Mem. Frame Capacity	Acc ↑	BWF ↓	Mem. Frame Capacity	Mem. Frame Capacity	Acc ↑	BWF ↓
EWC	10	None	None	5.81%	16.05%	None	None	4.02%	5.32%	None	None	9.51%	98.94%
MAS	10	None	None	7.81%	10.12%	None	None	8.11%	0.18%	None	None	10.89%	11.11%
EWC	20	None	None	2.95%	32.70%	None	None	1.28%	3.77%	None	None	4.71%	92.12%
MAS	20	None	None	4.25%	5.54%	None	None	4.61%	0.1%	None	None	5.90%	5.31%
Naive	10	8000	2×10^6	30.14%	41.30%	4000	15.5×10^6	47.20%	20.64%	2020	3.69×10^5	91.42%	7.43%
iCaRL	10	8000	2×10^6	32.04%	38.74%	4000	15.5×10^6	48.53%	19.72%	2020	3.69×10^5	80.97%	18.11%
BiC	10	8000	2×10^6	27.90%	51.96%	4000	15.5×10^6	51.96%	24.27%	2020	3.69×10^5	78.16%	18.49%
Naive	20	8000	2×10^6	23.47%	48.05%	4000	15.5×10^6	40.78%	23.18%	2020	3.69×10^5	87.40%	10.96%
iCaRL	20	8000	2×10^6	26.73%	42.25%	4000	15.5×10^6	43.33%	21.57%	2020	3.69×10^5	76.59%	21.83%
BiC	20	8000	2×10^6	23.06%	58.97%	4000	15.5×10^6	46.53%	15.95%	2020	3.69×10^5	70.69%	24.90%

Table 2. **Baseline Video CIL Results**. We report the average accuracy (Acc) and the backward forgetting (BWF) at 10 and 20 tasks from three action recognition benchmarks. Regularization-based methods (at the top of the table), under-perform memory-based methods, shown at the bottom. Consistent with results in the image domain, the longer 20-task sequences are more challenging for all methods across all dataset. We highlight that the reported improvements in the image domain do not directly translate to the video domain, as no single method obtains the best performance in every setup.

randomly samples instances for memory, outperforms the other baselines on UCF101, iCaRL and BiC are better on Kinetics and ActivityNet. We believe that fixing this discrepancy by finding memory-based methods designed specifically for video is an important research direction.

Number of Tasks and Forgetting. We observe that the longer (20-task) sequence is a more challenging setup, where the average accuracy is always lower for any method on any of the three datasets. Following this trend, all memory-based methods forget more as the number of tasks in the sequence increases. Similar to the image domain, evaluating on long sequences of tasks accentuates the shortcomings of CIL methods, and is thus suitable for the study of strategies that mitigate forgetting. We pose closing the forgetting gap between the 10-task and 20-task scenarios as an important research direction.

Relevance of Memory Size in Video. UCF101 stands out in Table 2, since memory-based methods perform surprisingly well in this datatset. In fact, the 91.42% accuracy obtained by the naive rehearsal baseline on the UCF101 10-task CIL is almost on par with the 94.9% accuracy reported for training TSN [35], which is the the backbone in our experiments, on all UCF101 classes simultaneously. This creates a sharp contrast with Kinetics and ActivityNet results, where the best CIL baseline achieve 32.04% and 51.96% respectively. In comparison, TSN trained on all the action classes of the whole dataset at once achieves 73.9% on Kinetics [40] and 88% on ActivityNet [35].

Park *et al*. [26] conducted experiments on UCF101 and made the observation that storing a subset of frames in memory results in a similar performance to storing the whole video [26]. Our hypothesis is that such good performance and apparent invariance to frame sampling is an artifact specific to the UCF101 dataset, and that it is not the norm in video class incremental learning. Our experiments on Kinetics and ActivityNet in Table 3 show that this is indeed not the case for more challenging video datasets. In particular for ActivityNet, storing four frames per video in memory, results in an overall decrease of accuracy by 27% compared to storing all frames. We hypothesize that the temporal dependencies in Kinetics and ActivityNet are more complex and more relevant for the task, thus models with naive sampling strategies struggle to remember older tasks from severely down-sampled videos.

4.2. Remembering from Down-sampled Videos.

To avoid incurring large memory requirements, it is favorable for video CIL models to down-sample videos before storing them for future replay. This means the model will learn a new task with full-resolution videos, but its memory will be composed of temporally sub-sampled versions of videos belonging to older tasks. Unfortunately, the distribution shift from the original training data to the modified stored exemplars causes a decline in the accuracy, which is evident from the Kinetics and ActivityNet results in Table 3.

To mitigate forgetting, continual video action recognition models must learn robust action embeddings, which are invariant to the temporal resolution of the video. We use our temporal consistency loss to train the CIL model jointly on both full-resolution videos and temporally down-sampled videos. Our proposed strategy helps achieve this desirable invariance. Table 3 reports the results of using our consistency loss, which is explained in Section 3.2, to mitigate forgetting in models that remember from down-sampled videos. Since no single memory-based method consistently outperforms the other methods across all datasets, we choose the more established baseline iCaRL [27] to perform the sub-sampled memory experiments.

Consistency Regularization on Kinetics. The last three rows of Table 3 summarize the effect of applying our temporal regularization on limited size memories. Temporal

Model	Frames per video	Kinetics			ActivityNet-Trim			UCF101		
		Mem. Frame Capacity	Acc ↑	BWF ↓	Frame Capacity	Acc ↑	BWF ↓	Mem. Frame Capacity	Acc ↑	BWF ↓
iCaRL	4	3.2×10^4	30.73%	40.36%	1.6×10^4	21.63%	36.98%	8.08×10^3	80.32%	17.13%
iCaRL	8	6.4×10^4	32.04%	38.48%	3.2×10^4	21.54%	33.41%	16.16×10^3	81.12%	18.25%
iCaRL	16	12.8×10^4	31.36%	38.74%	6.4×10^4	25.27%	29.71%	32.32×10^3	81.06%	18.23%
iCaRL	ALL	2×10^6	32.04%	38.74%	15.5×10^6	48.53%	19.72%	3.69×10^5	80.97%	18.11%
iCaRL+TC	4	3.2×10^4	35.32%	34.07%	1.6×10^4	42.99%	23.82%	8.08×10^3	73.85%	26.35%
iCaRL+TC	8	6.4×10^4	36.24%	33.83%	3.2×10^4	45.73%	18.90%	16.16×10^3	74.25%	25.27%
iCaRL+TC	16	12.8×10^4	36.54%	33.53%	6.4×10^4	44.04%	22.82%	32.32×10^3	75.84%	23.23%

Table 3. **Ablation study results with different memory sizes**. We compare iCaRL [27] with and without Temporal Consistency (iCaRL+TC) on the 10-task trimmed action recognition setups. iCaRL experiences a loss of accuracy as the memory size decreases in the challenging Kinetics and ActivityNet-Trim setups. Applying TC allows us to reduce the memory size by 2 orders of magnitude while retaining the performance in ActivityNet-Trim, and it even outperforms the version of iCaRL that uses all frames in Kinetics.

consistency regularization (labeled TC in the table) reduces forgetting on the 10-task Kinetics split regardless of how many frames per video are stored. In particular, our best baseline (iCaRL) tested on a memory of 16, 8, or 4 frames per video is significantly improved when our temporal consistency term is added to iCaRL's loss objective. For example, the accuracy is improved by more than 4.5% in the scenario where 4 frames per video are stored. It is worth highlighting that adding temporal consistency enables us to store as few as 4 frames per video and yet achieve more than 3% accuracy improvement over storing full videos, which requires about 100 times greater *memory frame capacity*.

Consistency Regularization on Trimmed ActivityNet. We observe a similar trend on ActivityNet with temporal consistency resulting in even more sizable improvements. Specifically, adding the regularization term with a model that has access to a memory consisting of 8 frames per video results in a massive 24% improvement. Moreover, with 4 frames per video stored in memory, our temporally consistent model comes close to achieving similar performance to the model that uses full-resolution videos. In fact, this significantly closes the 27% accuracy gap between storing full-resolution videos in memory and storing 4 frames per video when our regularization is not used. Our results show that our method is most relevant for datasets that require more sophisticated temporal reasoning like ActivityNet.

How Many Frames Should be Stored in Memory? A major challenge for adapting continual learning methods to video is that videos consume significantly more memory than images due to the added temporal dimension. Our experiments in Table 3 show that a consistency-based training framework makes it possible to down-sample videos before storing them for future replay, resulting in remarkably small memories and minor performance degradation. In particular on the challenging dataset Kinetics, iCaRL without TC trained with full-resolution memory achieves an average accuracy of 32.04%. Yet, adding the TC loss and training on memory of only 8 frames per video exemplar results in an even better average performance of 36.24%. This is impressive, since the average number of frames per video in Kinetics is 250, which translates to only 3.2% of the average Kinetics video being stored in memory.

Our experiments on ActivityNet-Trim also show a similar trend towards alleviating large memory requirements. Since 8 frames represent 0.21% of the average ActivityNet video, naively storing 8-frame exemplars with no temporal consistency in training results in a 27% drop in accuracy. Using consistency regularization and 8 frames per memory sample, we are able to reduce the difference in accuracy from 27% to only 3%.

Frame Sampling in UCF101. We revisit our assertion that remembering old tasks in UCF101 is unaffected by the number of frames stored in memory. We vary the number of frames per video used: from storing all frames to storing 16, 8, and 4 frames. The results reported in Table 3 clearly show that the performance of iCaRL is almost the same in all these different scenarios, validating the claim that UCF101 is not a prototypical dataset to evaluate CIL methods. For completeness, we also run the same set of experiments using the proposed consistency regularization scheme. We notice that it does not help increase the performance. This is unsurprising for two reasons. **(i)** Class incremental learning on UCF101 is already not that much more challenging than training on the whole dataset in one task, as we showed in Section 4.1. **(ii)** As the first half of Table 3 shows, class incremental learning with a memory buffer on UCF101 is invariant to the number of frames per video stored in memory. This can be explained by the strong scene bias exhibited in UCF101 [9]. Thus, incorporating the TC loss in this dataset, which has a very small accuracy gap between CIL training and fully-supervised training, is not expected to improve accuracy.

UCF101 for video CIL. In conclusion, we still recommend using the UCF101 splits for prototyping video CIL methods. However, due to its simplicity, we encourage the

Model	Frames per video	Mem. Frame Capacity	ActivityNet-Untrim		ActivityNet-Trim	
			Acc ↑	BWF ↓	Acc ↑	BWF ↓
iCaRL	4	1.6×10^4	16.28%	32.75%	21.63%	36.98%
iCaRL	8	3.2×10^4	16.67%	31.96%	21.54%	33.41%
iCaRL	16	6.4×10^4	21.27%	28.94%	25.27%	29.71%
iCaRL+TC	4	1.6×10^4	36.07%	22.39%	42.99%	23.82%
iCaRL+TC	8	3.2×10^4	40.29%	20.80%	**45.73%**	18.90%
iCaRL+TC	16	6.4×10^4	**40.45%**	21.21%	44.04%	22.82%

Table 4. **Ablation study results with trimmed and untrimmed videos**. All the experiments involve sequentially training on 10 tasks. ActivityNet-Untrim provides a more realistic and challenging setup to evaluate CIL models. We impose the strict resource constraint of 4, 8, and 16 frames per video stored in memory. Our temporal consistency approach significantly improves the accuracy of the baseline method trained on both ActivityNet setups with limited memory. The best performing setting for each data split is highlighted in the table.

community to evaluate new video CIL methods on the more challenging splits from Kinetics and ActivityNet.

Class Incremental Learning from Untrimmed Videos. As observed in Table 4, we perform a set of experiments to evaluate the realistic class incremental learning scenario with untrimmed videos and make a few interesting observations. First, ActivityNet-Untrim is more challenging than ActivityNet-Trim. iCaRL baseline [27] achieves a better performance on ActivityNet-Trim regardless of the number of frames per video stored in memory. Second, our temporal consistency regularization improves [27] by large margins in both ActivityNet setups. The progressive performance of iCaRL on the trimmed and untrimmed setups in the 10-task case, where 8 frames are used to represent a memory video, are plotted in Figure 3. Our regularization loss enhances the performance by 24% in both cases.

Why Does Consistency Regularization Work? We hypothesize that the large accuracy gains garnered by adding the consistency regularization loss are due to two reasons. First, regularization may make the training more stable. When training the model on augmented examples, we expect the model to learn more robust representations that can make learning new tasks easier. Second, consistency regularization forces the backbone to learn action embeddings that are invariant to the number of frames used to represent the video. Current video CIL methods store down-sampled videos in the episodic memory. However, without applying the consistency loss, the model naturally learns completely different features for densely sampled videos and sparsely sampled videos. Thus, the model struggles to remember the old action representations from this memory of down-sampled videos. This distribution shift between videos of different temporal resolution is especially apparent in datasets with more complicated temporal dependencies like ActivityNet. This points to why our approach improves the performance by large margins on this dataset and suggests TC loss does not skew the models towards spurious scene features, rather it manages to retain meaningful temporal features.

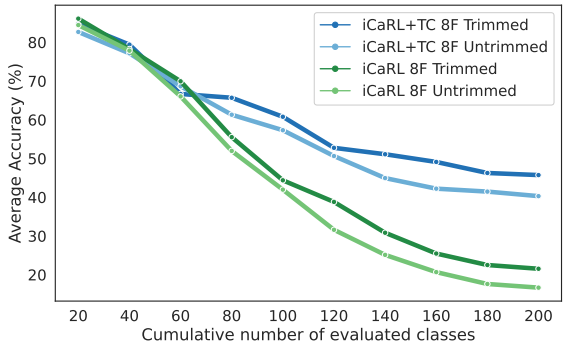

Figure 3. **Average Accuracy in the Validation Set.** We train iCaRL sequentially on 10 tasks and enforce a memory constraint of 8 frames per video. Our temporal consistency (TC) loss significantly improves iCaRL's performance on ActivityNet-Trim and ActivityNet-Untrim

5. Conclusion and Limitations

In this paper, we propose and analyze vCLIMB, a continual learning benchmark for video action recognition. We expose and tackle the unstudied accuracy drop, which is experienced by memory-based video CIL models and is caused by frame sub-sampling. In our experiments, we sample frames uniformly and leverage a consistency loss to significantly alleviate that accuracy drop, by up to 24% in CIL untrimmed video classification. We think that exploring non-uniform sampling strategies is another promising direction, but we leave that exploration for future work.

Acknowledgments. This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-CRG2021-4648. Authors also thank Centro Nacional de Inteligencia Artificial CENIA, FB210017, BASAL, ANID.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Eur. Conf. Comput. Vis.*, September 2018. 2, 4, 5

[2] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Adv. Neural Inform. Process. Syst.*, pages 5050–5060, 2019. 2

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Eur. Conf. Comput. Vis.*, pages 850–865. Springer, 2016. 1

[4] Jessica Bursztynsky. Tiktok says 1 billion people use the app each month, 2021. 1

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 961–970, 2015. 1, 3, 4, 5

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6299–6308, 2017. 1, 3, 5

[7] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9233–9242, 2020. 2

[8] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Eur. Conf. Comput. Vis.*, pages 532–547, 2018. 2

[9] Jinwoo Choi, Chen Gao, C. E. Joseph Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Adv. Neural Inform. Process. Syst.*, 2019. 7

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255. Ieee, 2009. 1

[11] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5138–5146, 2019. 1

[12] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Int. Conf. Machine learning*, pages 647–655. PMLR, 2014. 1

[13] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 1

[14] Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. In *NeurIPS Continual learning Workshop*, 2018. 5

[15] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563*, 2017. 2

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[17] 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, 4, 5

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.*, 25:1097–1105, 2012. 1

[19] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Int. Conf. Learn. Represent.*, 2017. 2

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014. 1

[21] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12245–12254, 2020. 2

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3431–3440, 2015. 1

[23] David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Adv. Neural Inform. Process. Syst.*, volume 30. Curran Associates, Inc., 2017. 3, 4

[24] Jiawei Ma, Xiaoyu Tao, Jianxing Ma, Xiaopeng Hong, and Yihong Gong. Class incremental learning for video action classification. In *IEEE Int. Conf. Image Process.*, pages 504–508. IEEE, 2021. 1, 3

[25] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11321–11329, 2019. 2

[26] Jaeyoo Park, Minsoo Kang, and Bohyung Han. Class-incremental learning for action recognition in videos. In *Int. Conf. Comput. Vis.*, pages 13698–13707, 2021. 1, 2, 3, 6

[27] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2001–2010, 2017. 2, 4, 5, 6, 7, 8

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inform. Process. Syst.*, 28:91–99, 2015. 1

[29] Amelie Royer and Christoph H Lampert. Classifier adaptation at prediction time. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1401–1409, 2015. 1

[30] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Adv. Neural Inform. Process. Syst.*, pages 1163–1171, 2016. 2

[31] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *arXiv preprint arXiv:1705.08690*, 2017. 2

[32] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 3, 5

[34] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *AAAI Conf. Artificial Intelligence Worksh.*, 2012. 1

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2740–2755, 2019. 5, 6

[36] Susan Wojcicki. Youtube at 15: My personal journey and the road ahead, 2020. 1

[37] Maciej Wołczyk, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. In *Adv. Neural Inform. Process. Syst.*, 2021. 3

[38] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2019. 2, 4, 5

[39] Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[40] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Eur. Conf. Comput. Vis.*, pages 318–335, 2018. 6

[41] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Adv. Neural Inform. Process. Syst.*, volume 27. Curran Associates, Inc., 2014. 1

[42] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Int. Conf. Machine learning*, pages 3987–3995. PMLR, 2017. 2

[43] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *Int. Conf. Learn. Represent.*, 2020. 2

[44] Hanbin Zhao, Xin Qin, Shihao Su, Zibo Lin, and Xi Li. When video classification meets incremental classes. *arXiv preprint arXiv:2106.15827*, 2021. 1, 2, 3

[45] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Adv. Neural Inform. Process. Syst.*, 2020. 2