Supplementary Materials of 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

1. Supplementary

We complement the empirical evaluations of section 4, and present additional in-depth analysis of the results. First we assess the Average Accuracy and the BWF individually along the full set of tasks. We conduct this analysis for Kinetics, ActivityNet-Trim and ActivityNet-Untrim. These results highlight the effectiveness of the Temporal Consistency loss as the class incremental learning tasks progress.

1.1. The Naive Memory-based Baseline

Our naive memory-based baseline method is a simpler version of iCaRL. It follows iCaRL's inference approach, which is based on metric learning. The difference between the two methods, is the way they select the videos to store for future replay. Our naive baseline selects the videos randomly, while iCaRL estimates how close each video sample is to its class prototype. Thus, iCaRL is expected to perform better than the naive baseline. Experimentally, we find that the naive baseline outperforms iCaRL on UCF101. However, iCaRL achieves better results on the more challenging splits drawn from Kinetics and ActivityNet.

1.2. Temporal Progression of Temporal Consistency (TC) Loss

Figure 1 shows the Average Accuracy of the baseline methods as they sequentially learn more classes in vCLIMB. Not only does the use of Temporal Consistency (TC) reduce the required memory size considerably, but it also keeps the performance almost at the same level of methods that use full memory. In particular for ActivityNet-Trim, 1a shows that most of the performance is retained using only 8 frames per video in memory when using TC. Such an improvement is notable for a dataset like ActivityNet-Trim, which is composed of long video sequences. Notably for kinetics 1b, TC allows for storing only 8 frames while even improving the performance.

Moreover, our TC loss achieves outstanding results on ActivityNet-Untrim and ActivityNet-Trim with the Naive memory-based method, as shown in Table 1. It improves the performance by 21% on both benchmarks when 4 and 8 frames per video are saved. This shows the capacity of our TC loss to be used with different memory-based methods.

1.3. Sensitivity to the number of stored frames

As Figures 2 and 3 show, in less-controlled scenarios like ActivityNet-Trim, which have long videos, the Naive and iCaRL methods are more susceptible to the number of stored frames per video than when they deal with short videos like in Kinetics. It is precisely in these scenarios where our TC loss is essential to reduce the memory consumption without affecting the model performance. In this scenario, we reduce approximately 99.79% in memory usage leveraging the TC loss.

1.4. What Actions Require Temporal Consistency (TC) to Learn in the Class Incremental Setup of ActivityNet-Trim?

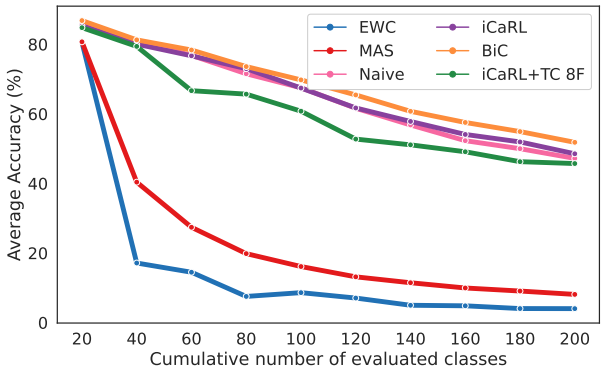
We analyze the per-class performance of iCaRL in the class incremental learning scenario, with the constraint of storing 8 frames per video in memory. We train two iCaRL models, one trained with TC and the other without TC, sequentially on 10 tasks of ActivityNet-Trim. Figure 4 visualizes the performance of the two models on a sample of challenging classes, ordered by the performance of the TC model. Using our TC approach consistently and significantly improves the performance on most of the sampled classes. For some classes that require more temporal and motion reasoning, like playing kickball and hula hoop, the model without TC completely fails. Incorporating the TC loss improves the accuracy by large margins for these

*Work done during an internship at KAUST.

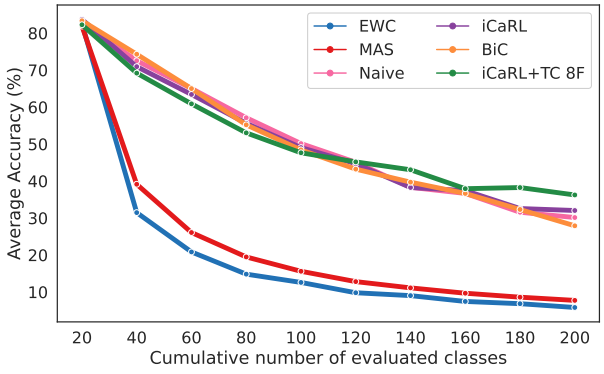
Model	Frames per video	Mem. Frame Capacity	ActivityNet-Untrim		ActivityNet-Trim	
			Acc ↑	BWF ↓	Acc ↑	BWF ↓
Naive	4	1.6×10^4	15.67%	32.97%	22.27%	35.24%
Naive	8	3.2×10^4	18.25%	31.71%	22.44%	36.06%
Naive+TC	4	1.6×10^4	37.58%	23.93%	43.11%	23.58%
Naive+TC	8	3.2×10^4	39.80%	19.59%	43.97%	20.67%

Table 1. **Ablation study results with trimmed and untrimmed videos**. This table complements Table 4 of the main paper. 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 and 8 frames per video stored in memory. Results were shown in the paper for the iCaRL baseline. Here, we present the results for the Naive baseline, which selects memory samples randomly. Our temporal consistency approach improves the accuracy of both baseline methods trained with limited memory on ActivityNet-Trim and ActivityNet-Untrim by large margins. For Naive trained on 4 frames per memory video, TC results in a 22% improvement.

actions. Similarly, while iCaRL without TC achieves an accuracy of less than 8% on the action of playing squash, training iCaRL with the TC loss results in an an accuracy of around 68% on playing squash, enhancing the performance on that category by 60%.

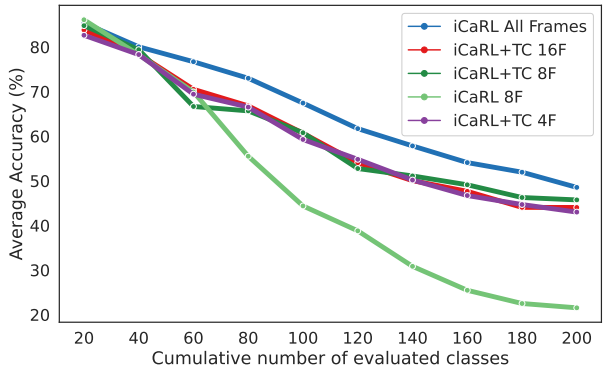


(a)

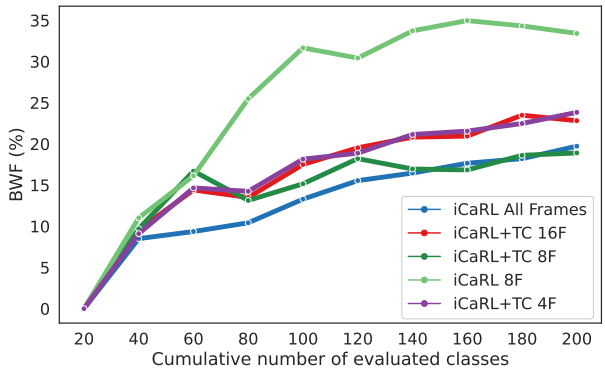


(b)

Figure 1. **Average Accuracy in the Validation and Test Sets of ActivityNet-Trim and Kinetics.** We sequentially train different memory-based methods (iCaRL, BiC, iCaRL+TC) and regularization-based methods (EWC, MAS) on 10 tasks. Our temporal consistency (TC) loss achieves competitive results on: (a). ActivityNet-Trim and (b). Kinetics while using significantly smaller memory. Specifically, it stores around 100 times fewer frames than the other memory-based methods evaluated on Kinetics but still outperforms them.

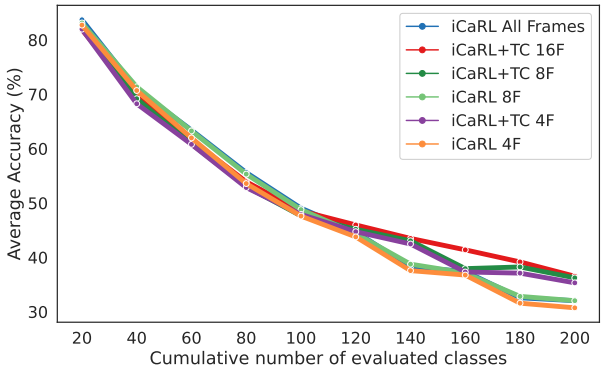


(a)

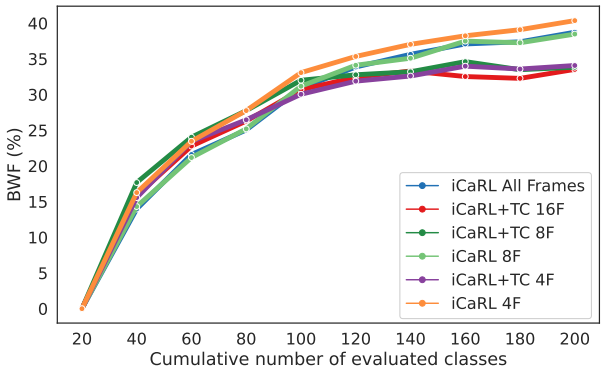


(b)

Figure 2. **Consistency Regularization Improves the Performance in the Validation Set of ActivityNet-Trim.** iCaRL experiences a sharp increase in backward forgetting when trained with memories of severely down-sampled videos (an example is shown here with 8 frames, but more examples are reported in Table 4 of the paper.) Our temporal consistency (TC) approach significantly alleviates this forgetting. Storing only 8 frames per video in memory, iCaRL trained with TC reports similar average accuracy and backward forgetting to the baseline iCaRL, which uses full-resolution videos for memory replay.



(a)



(b)

Figure 3. **Consistency Regularization Improves the Performance in the Test Set of Kinetics.** iCaRL experiences the same increase in backward forgetting when trained with memories of severely down-sampled videos on Kinetics (an example is shown here with 8 frames, but more examples are reported in Table 3 of the paper.) Our temporal consistency (TC) approach significantly alleviates this forgetting. Storing only 8 frames per video in memory, iCaRL trained with TC reports even higher average accuracy than the baseline iCaRL, which uses full-resolution videos for memory replay.

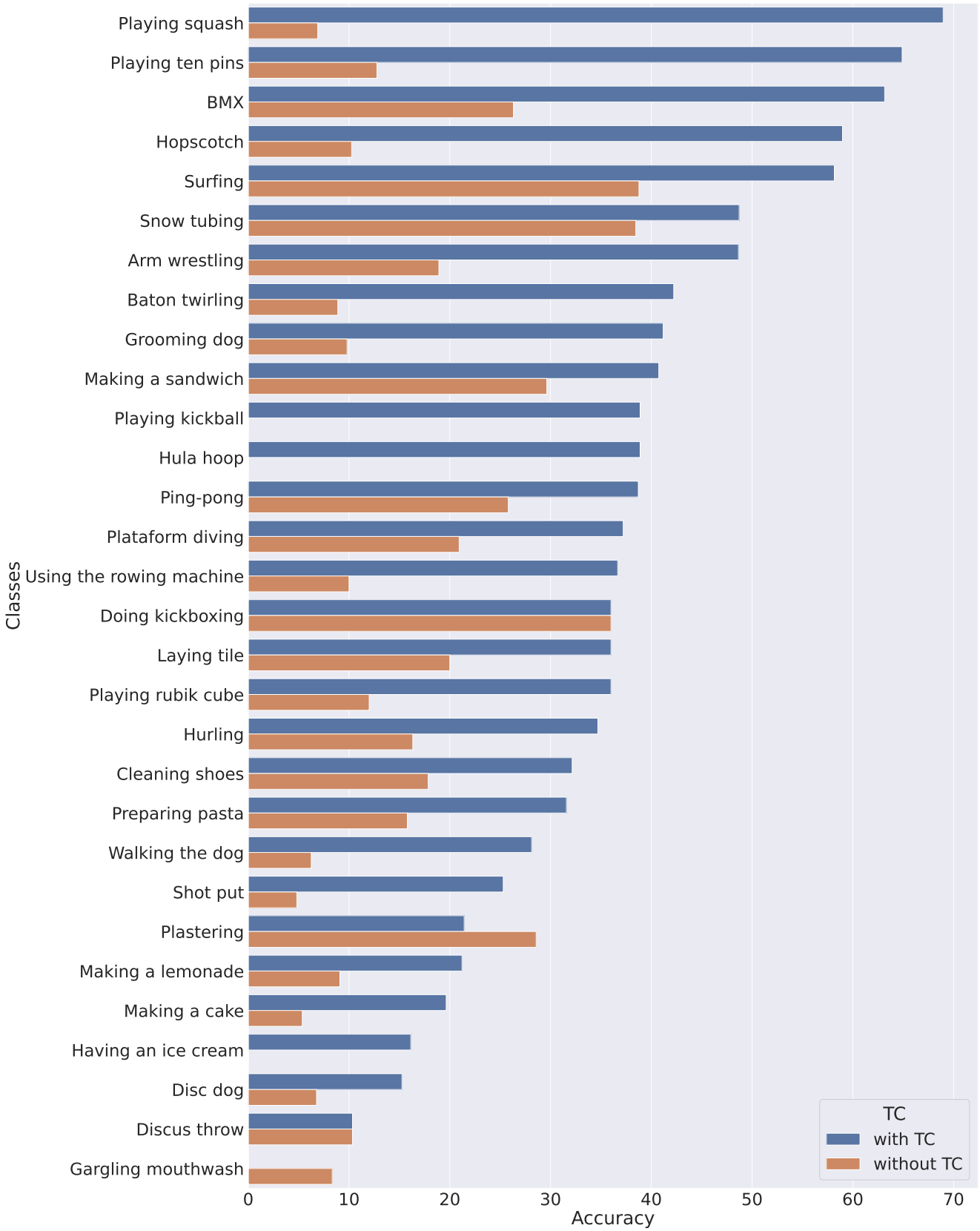


Figure 4. **Temporal Consistency Training Improves iCaRL's Ability to Recognize Challenging Actions.** A more temporally consistent model is better able to recognize actions with fast movements, like playing squash and ten pins (bowling.)