Supplementary Material:
Self-supervised Video Representation Learning with Cascade Positive Retrieval

Cheng-En Wu[1]*, Farley Lai[2], Yu Hen Hu[1], Asim Kadav[2]
[1] Department of Electrical and Computer Engineering, University of Wisconsin-Madison, WI, USA
[2] NEC Laboratories America, Inc., San Jose, CA, USA
{cwu356, yhhu}@wisc.edu
farleylai@icloud.com asimkadav@gmail.com

1. Implementation Details

1.1. Self-supervised Pretraining

We use MoCo [2] as the contrastive learning framework and S3D [3] as the feature extractor to implement CPR. Note that MoCo is not required but is useful to save memory usage. Therefore, it is simply a coincidence that the baseline CoCLR [1] uses MoCo and we have applied CPR to other work not using MoCo such as IIC as well. At pretraining stage, two fully connected layers (FC1024→ReLU→FC128) are used as a projection head after the global average pooling layer to obtain the embedding features but the projection head is removed for the model to perform downstream tasks. Following CoCLR we set the momentum to 0.999, the temperature to 0.07, and the size of the queue to 2048 on every dataset Training each model on UCF 101 , we use ADAM as our optimizer with an initial learning rate of 10^{-3} and weight decay of 10^{-5}, where the learning rate is multiplied by 0.1 at 300 and 350 epochs.

1.2. Action Recognition

For action recognition task, we use ADAM to optimize the model for 500 epochs with a batch size of 16 on two GPUs. The initial learning rate is set to 10^{-3}, where the learning rate is decayed by 0.1 at 400 and 450 epoch respectively. The momentum is 0.9 and the weight decay is 10^{-3}. At evaluation stage, we follow the practice of CoCLR to uniformly sample 32 frames from each video, perform ten-crop to 128×128 pixels, and then average their predictions to become the final video prediction.

2. Additional Results

2.1. Class Mining Recall (CMR)

To evaluate the overall mining quality, we further define Class Mining Recall (CMR) in Eq 1 to measure how a model is able to successfully mine distinct true positives from a certain class in one training epoch.

$$Class\ Mining\ Recall = \frac{\#Distinct\ TP\ Selected}{\#Total\ Class\ Instances} \quad (1)$$

As shown in Figure 1, we present the full CMR in the UCF101 classes. Closer inspection of the figure reveals that CPR has the Top-3 classes are *Diving* (100%) , *PommelHorse* (100%), and *UnevenBars* (100%). On the other hand, the bottom 3 classes are *Lunges* (27.8%), *Haircut* (33.3%), and *HandstandWalking* (33.3%). Furthermore, we demonstrate video frames from these classes above to visualize their content including human action and background.

In the overall evaluations of all 101 classes, CPR scores higher CMR than baseline CoCLR in 48 classes while the baseline mines better only in 20 classes. It is even in the rest classes. Regarding the classes that the baseline has higher CMR, it may that those positive classes highly correlate with a single view while CPR sometimes does not help much after visual inspection. A further comparison of the number of 100% CMR between both methods, CPR obtains eight perfect CMRs across 101 classes, which exceeds two classes compared to the baseline getting six. In addition, we evaluate the entire performance of CPR by applying median CMR. Our approach achieves the median CMR of **83.3%** across all the classes, which outperforms the baseline with the median CMR of 77.8%. This outcome shows a great improvement with a margin of **5.5%**. In summary, these empirical evidence validates the effectiveness of our approach in mining the higher quality positives than the baseline.

[*] Work done as a NEC Labs intern in 2021.

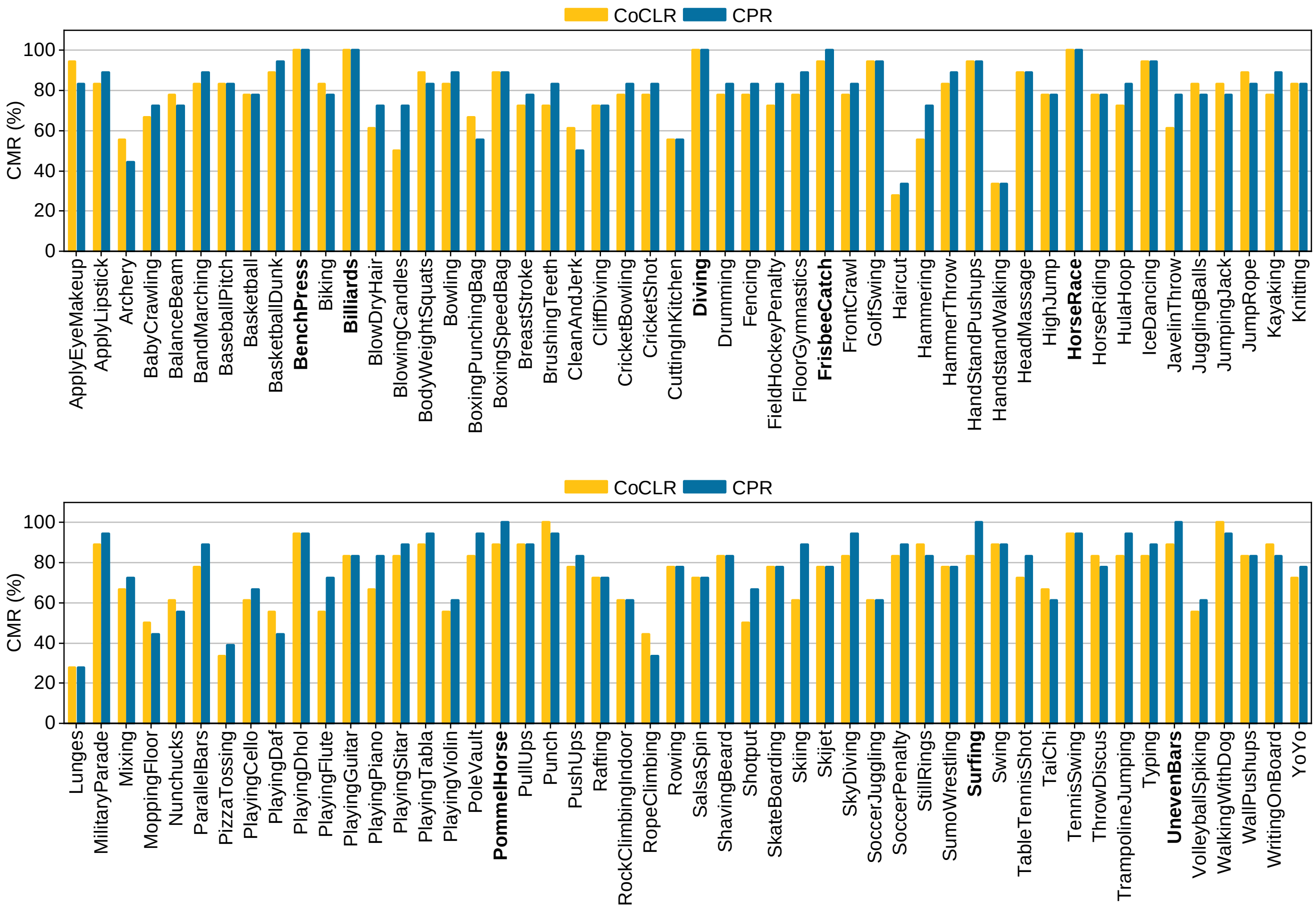

Figure 1. Class Mining Recall (CMR) per action class on UCF101. There are 101 action classes which are listed in alphabetical order. The upper bar chart covers the first 50 action classes while the lower bar chart covers the rest of the action classes. Eight action classes appeared in bold font represent cases where CPR achieves 100% CMR. They are 1. *BenchPress*, 2. *Billiards*, 3. *Diving*, 4. *HorseRace*, 5. *FrisbeeCatch*, 6. *PommelHorse*, 7. *UnevenBars*, and 8. *Surfing*. Note that we measure the CMR of both approaches in their last training epoch.

References

[1] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *NeurIPS*, Oct. 2020. 1

[2] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, Nov. 2019. 1

[3] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 1