Frame-wise Action Representations for Long Videos via Sequence Contrastive Learning - Supplementary Material -

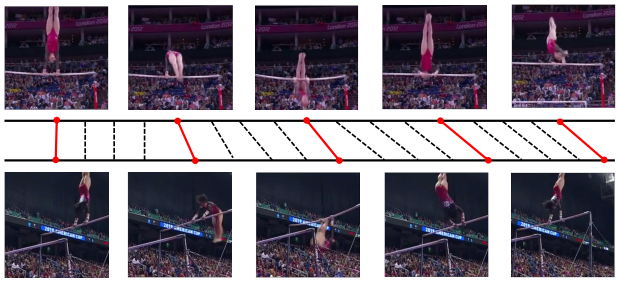
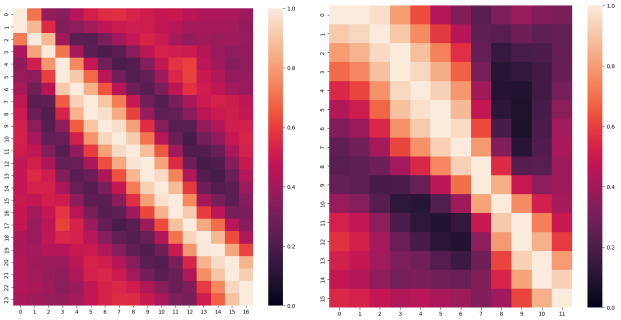


Figure 1. Visualization of video alignment on FineGym dataset. Please refer to video demos in our supplementary materials for more visualization results.



(a) Pouring dataset. (b) PennAction dataset.

Figure 2. We randomly select two videos recording the same process (or action) from Pouring (or PennAction) dataset and compute the similarity matrix for frame-wise representations extracted by our method. The similarities are normalized for better visualization.

A. More Results

In this section, we show visualization results of video alignment and fine-grained frame retrieval.

A.1. Video Alignment

Given two videos recording the similar action or process, the goal of video alignment is to find the temporal correspondence between them. Firstly, we use our framework to extract the frame-wise representations for two randomly selected videos. Then we compute the cosine similarities between the frame-wise representations of two videos and utilize the famous dynamic time warping (DTW) algorithm on the similarity matrix to find the best temporal alignment. Figure 1 shows an example from FineGym test set. Please refer to video demos in our supplementary materials for more visualization results.

We also randomly select two videos recording the same process (or action) from Pouring (or PennAction) dataset, and similarly, we can compute the similarity matrix which is rendered as a heatmap in Figure 2. We observe that the diagonal is highlighted, which means our approach find the favorable alignment between two correlated videos. We also give video demos in our supplementary materials.

Query Top-5 Retrieved Frames

Figure 3. Visualization of fine-grained frame retrieval on FineGym datast by using our method.

A.2. Fine-grained Frame Retrieval

In Figure 3, we present the visualization results of fine-grained frame retrieval on FineGym dataset. To be specific, we feed the video containing the query frames into our CARL framework to generate query features, and simi-

Method	0.5	0.75	0.95	Average
G-TAD w. 2stream	50.36	34.60	9.02	34.09
G-TAD w. ours	51.22	35.19	8.54	34.46

Table 1. Temporal action localization on ActivityNet v1.3.

Method	Classification	Progress	τ
Contrastive baseline	88.05	0.898	0.891
SCL (ours)	**93.07**	**0.918**	**0.985**

Table 2. Compare our SCL with contrastive baseline, which uses the corresponding frame in the other view as the positive sample.

Training Dataset	Classification	Progress	τ
K400	91.9	0.903	0.949
K400 $\rightarrow$ PennAction	**93.9**	**0.908**	**0.977**

Table 3. Our CARL pre-trained on Kinetics-400 shows outstanding transfer ability on PennAction. Fine-tuning the pre-trained model on PennAction further boosts the performance.

larly, we can extract frame-wise features for the rest videos in the test set. We simply compute the cosine similarity between query features and frame-wise features from candidate videos to obtain top-5 retrieved frames as shown in Figure 3. The retrieved frames have similar semantics with the query frame, though the appearances, the camera views, and the backgrounds are different, which suggests our method is robust to these factors.

A.3. Action Localization

To show the potential of our method on large datasets and more downstream tasks, we optimize the frame-wise features via our self-supervised method on ActivityNet [2]. Then we use G-TAD [4] on the top of the features (without fine-tuning) to perform temporal action localization. As shown in Table 1, we use mAP(%) at {0.5, 0.75, 0.95} tIoU thresholds and the average mAP across 10 tIoU levels for evaluation. In contrast to the supervised two-stream model [3], our method does not need any video labels while achieving better performance.

A.4. Compare with Contrastive Baseline

We compare our SCL with the contrastive baseline which only uses the corresponding frame in the other view as the positive sample and ignores temporal adjacent frames. As Table 2 shows, our SCL can more efficiently employ the sequential information and thus achieves better performance.

A.5. Kinetics-400 Pre-training

To show our method can benefit from large-scale datasets without any labels, we train our CARL on Kinetics-400 [1]. As Table 3 shows, the frame-wise representations trained on Kinetics-400 shows outstanding generalization on PennAction dataset. Moreover, fine-tuning the pre-trained model on PennAction by using our CARL further boosts the performance, e.g., + 2% classification improvement.

References

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[3] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[4] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, 2020.