–Supplementary Material–
Unsupervised Pre-training for Temporal Action Localization Tasks

Can Zhang[1] Tianyu Yang[2] Junwu Weng[2] Meng Cao[1] Jue Wang[2] Yuexian Zou[1]✉
[1]School of Electronic and Computer Engineering, Peking University [2]Tencent AI Lab
zhangcan@pku.edu.cn tianyu-yang@outlook.com WE0001WU@e.ntu.edu.sg
mengcao@pku.edu.cn arphid@gmail.com zouyx@pku.edu.cn

In this material, we provide more experimental details (Sec. 1), extended ablation studies (Sec. 2) and more qualitative visualizations (Sec. 3).

1. Details of Feature Extraction

In Sec. 4.2 of our paper, we give comparative results of our PAL with other state-of-the-art pre-training approaches on three challenging TAL tasks. For our implemented methods (MoCo-v2, TAC, PAL), we sample 4 frames (temporal resolution) with a spatial resolution of 112×112 for each clip. For those self-supervised methods designed for TAC tasks [1, 2, 4, 5], we borrow the pre-trained weights from their official implementations (CoCLR[1], XDC[2], VideoMoCo[3], RSPNet[4]) and directly use the pre-trained models to extract the features on the target TAL datasets. Note that the input resolutions (temporal & spatial) of these methods are dissimilar with each other. Considering that the inconsistent setting between training and testing will inevitably harm the performance, we follow their original input sizes but adjust the sampling frame rate and stride to ensure that the number of features extracted for each video is identical. Thereby, the reported FLOPs/clip can reflect the extraction efficiency of each method.

2. Extended Ablation Studies

In this section, we conduct extended ablation studies on our proposed PAL. For ease of experimentation, all the ablation studies are conducted with 100 training epochs on K400 and evaluated on the TAD task (ActivityNet v1.3 dataset).

2.1. Different clip numbers

In our experiment, we process $J = 8$ clips for each composited video. Here, we are interested in how the clip number affects the final performance. To ensure that there is at least one background clip on each side of the pseudo action region, the max clip length of the action region is limited to $J - 2$. The evaluation results under different input clip numbers (J) are listed in Table 1. The per-video FLOPs are also computed. As expected, more input clips (denser sampling) bring higher performances. The relative gain becomes smaller when the clip number is greater than 8, so we choose 8 as our default setting due to its best trade-off between effectiveness and efficiency.

Table 1. **Ablation study on different clip numbers.**

Clip Num. (J)	4	6	8	10	12
FLOPs/video	14.4G	21.4G	28.6G	35.8G	43G
mAP@AVG	31.2	32.2	32.8	33.0	33.1

3. Extended Feature Visualizations

In this section, more visualizations of the clip feature similarities are presented. Recall that in order to confirm the time-equivariant ability, we apply random temporal transformations on the action instances in real-world videos and investigate whether these changes will be reflected in feature space accordingly. For the temporal transformation, we consider two aspects: (1) *location* (moving forward *vs.* moving backward along time axis); (2) *scale* (down-sampling *vs.* up-sampling the action instances). We try different combinations of these two transformations. The extended visualization results are shown in Fig 1 to Fig. 4. As shown, compared to [3] and [5] features, our learned clip features are more similar within action regions and more different from the background regions in its temporal surroundings. Thus, our PAL features are more informative for TAL tasks. These visualizations also confirm our key idea that pseudo action localization (PAL) can enhance the temporal localization power of the feature encoder.

[1] https://github.com/TengdaHan/CoCLR
[2] https://github.com/HumamAlwassel/XDC
[3] https://github.com/tinapan-pt/VideoMoCo
[4] https://github.com/PeihaoChen/RSPNet

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[2] Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. Rspnet: Relative speed perception for unsupervised video representation learning. In *AAAI Conference on Artificial Intelligence*, volume 1, 2021. 1

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 3

[4] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 1

[5] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 1, 3

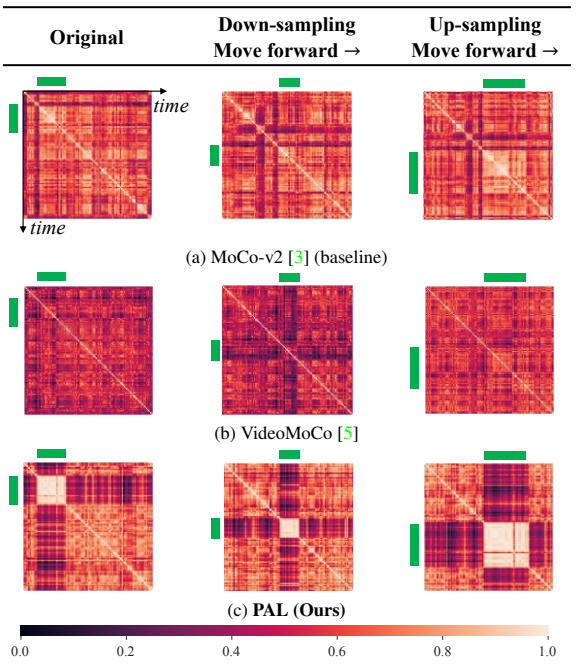

Figure 1. **Feature similarity visualization under different temporal transformations (2^{nd} & 3^{rd} columns) of ground-truth action instance.** The green bars represent the temporal extent of ground truth actions. Brighter color means higher similarity.

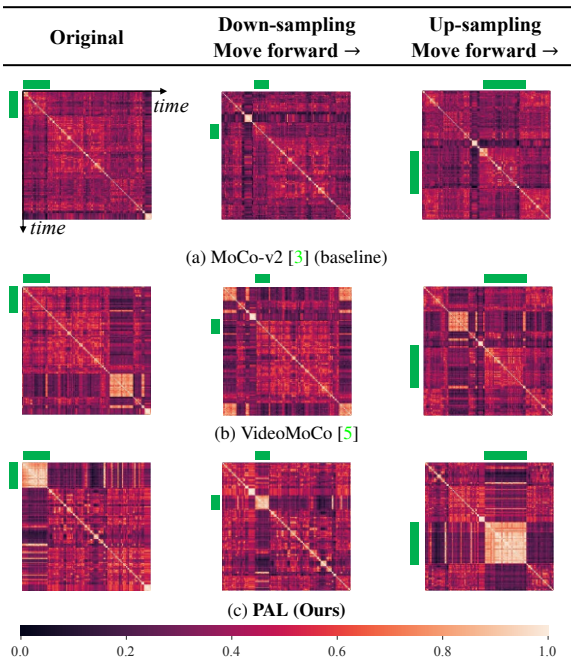

Figure 3. **Feature similarity visualization under different temporal transformations (2^{nd} & 3^{rd} columns) of ground-truth action instance.** The green bars represent the temporal extent of ground truth actions. Brighter color means higher similarity.

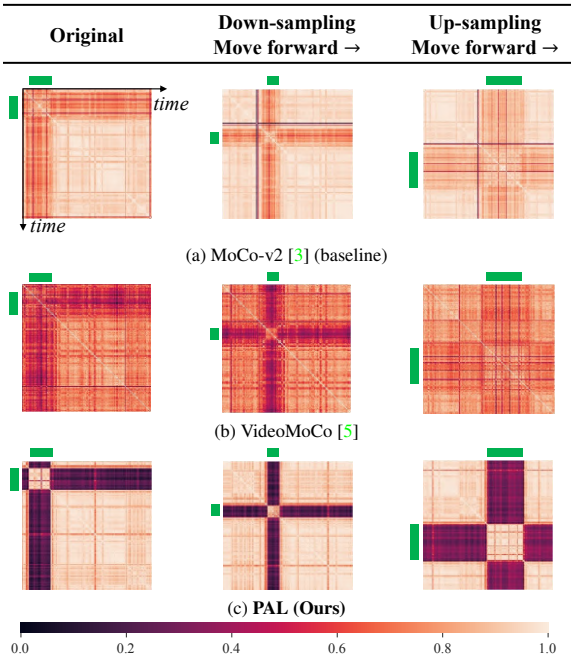

Figure 2. **Feature similarity visualization under different temporal transformations (2^{nd} & 3^{rd} columns) of ground-truth action instance.** The green bars represent the temporal extent of ground truth actions. Brighter color means higher similarity.

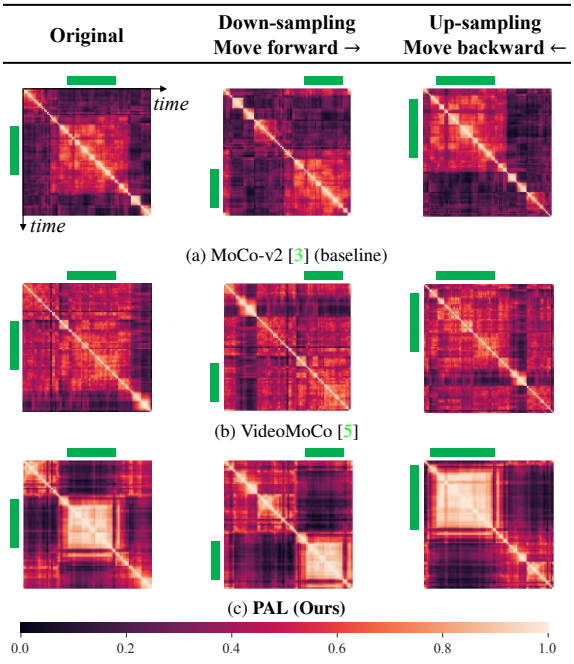

Figure 4. **Feature similarity visualization under different temporal transformations (2^{nd} & 3^{rd} columns) of ground-truth action instance.** The green bars represent the temporal extent of ground truth actions. Brighter color means higher similarity.