Supplementary Materials for "TCVM: Temporal Contrasting Video Montage Framework for Self-supervised Video Representation Learning"

Fengrui Tian[†1], Jiawei Fan[†2], Xie Yu[2], Shaoyi Du[1(✉)], Meina Song[2], and Yu Zhao[3]

[1] Xi'an Jiaotong University, Xi'an, China
[2] Beijing University of Posts and Telecommunications, Beijing, China
[3] Harbin Institute of Technology, Harbin, China
tianfr@stu.xjtu.edu.cn, {jwfan, yuxie130, mnsong}@bupt.edu.cn
dushaoyi@gmail.com, zhaoyu_zzz96@163.com

1 Action Classification Results on UCF101 and HMDB51 Datasets

In the paper, we have discussed that finetuning on the UCF101[15] and HMDB51[11] datasets are easy to overfit since these datasets are relatively small. We present the action classification results here for comparisons with other state-of-the-art methods in Table 1. The 30-view evaluation results are reported following [5] for R3D50 backbone. We sample 16 frames of each video and the input size is set as 16×224×224. The proposed method outperforms other state-of-the-art methods except MoDist[17], and STiCA[13]. MoDist uses the optical flow for self-supervised training. STiCA uses the multi-modality to improve cross-modal video representation learning. In contrast, the proposed method uses the RGB image only.

2 Ablation Study on OFC module

The effects of TC and VM modules are concluded as: **Table 2 demonstrates TC and VM are not independent in TCVM framework.** TC improves classification performance by balancing the effect between origin features and feature differences, while VM strengthens retrieval performance by dismissing scene bias. **(2) Table 3(a) of the main paper demonstrates OFC is necessary for retrieval tasks but less important for finetuning tasks.** Adding the OFC module enhances the retrieval performance (+24.6% on UCF101), while slightly improving (only 0.8%) the classification task.

3 Action Retrieval Visualizations

Fig. 1 shows four action retrieval visualization results on UCF101 and HMDB51 datasets. MoCo[8] is used to compare. Column 1 shows the Query actions.

Method	Year	Pretraining Dataset	Backbone	Input Size	UCF101	HMDB51
Huang[9]	2021	K400	R21D26	112	85.7	54.0
PacePred[16]	2020	K400	R21D	112	77.1	36.6
VideoPlayback[18]	2020	UCF101	R3D	112	66.5	29.7
VideoMoCo[12]	2021	K400	R3D	112	74.1	43.6
Mem-DPC[7]	2020	K400	R2D3D	224	78.1	41.2
MoSI[10]	2021	K400	R2D3D	224	70.7	48.6
SpeedNet[1]	2020	K400	S3D-G	224	81.1	48.8
DPC[6]	2019	K400	R3D34	224	75.7	35.7
STiCA[13]	2021	K400	R21D18	112	93.1	67.0
CoCon[14]	2021	K400	R3D34	224	79.1	48.5
MoDist[17]	2021	K400	R3D50	224	91.5	67.4
RSPNet[2]	2021	K400	R3D18	112	74.3	41.8
TCLR[3]	2021	K400	R3D18	112	84.1	53.6
Ding[4]	2021	K400	R21D	112	84.8	53.5
Ours†	2022	k400	R3D50	224	86.9	61.9

Table 1: Downstream action classification results on UCF101 and HMDB51 datasets. † denotes the results from 10×3 view evaluation following[5].

Method			Dataset		
OFC	TC	VM	UCF101	HMDB51	SSv2
✓			45.1	20.2	55.2
✓	✓		28.2	14.0	58.3
✓		✓	**56.7**	**26.2**	56.9
✓	✓	✓	56.1	25.6	**58.5**

Table 2: Ablation studies on OFC module.

Columns 2-4 present the action retrieval results by using MoCo. Columes 5-7 show the results by using the proposed method. Row 1 is the result from UCF101 dataset. Rows 2-4 are the results from HMDB51 dataset. It can be seen that our method reaches more accurate results in action retrieval task.

4 Pseudo Code

We present the pseudo code in the "TCVM_pseudo_code.py" file.

References

1. Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T.: Speednet: Learning the speediness in videos. In: CVPR. pp. 9922–9931 (2020)
2. Chen, P., Huang, D., He, D., Long, X., Zeng, R., Wen, S., Tan, M., Gan, C.: Rspnet: Relative speed perception for unsupervised video representation learning. In: AAAI. vol. 1 (2021)

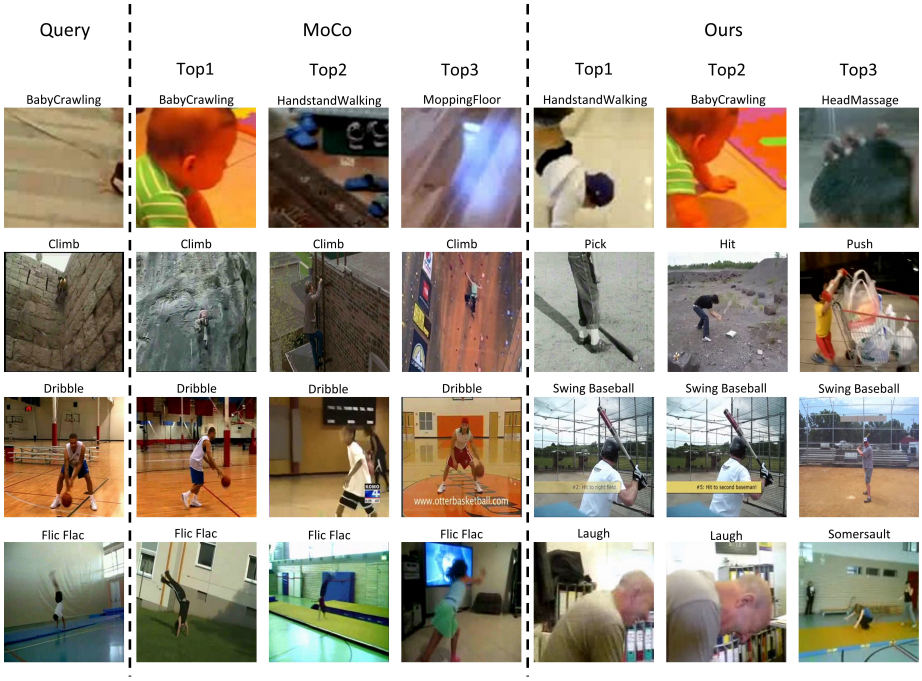

Fig. 1: Action Retrieval Visualizations on UCF101 and HMDB51 datasets. MoCo[8] is used to compare.

3. Dave, I., Gupta, R., Rizve, M.N., Shah, M.: Tclr: Temporal contrastive learning for video representation. arXiv preprint arXiv:2101.07974 (2021)
4. Ding, S., Li, M., Yang, T., Qian, R., Xu, H., Chen, Q., Wang, J.: Motion-aware self-supervised video representation learning via foreground-background merging. arXiv preprint arXiv:2109.15130 (2021)
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV. pp. 6202–6211 (2019)
6. Han, T., Xie, W., Zisserman, A.: Video representation learning by dense predictive coding. In: ICCV Workshops. pp. 0–0 (2019)
7. Han, T., Xie, W., Zisserman, A.: Memory-augmented dense predictive coding for video representation learning. In: ECCV. pp. 312–329 (2020)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
9. Huang, L., Liu, Y., Wang, B., Pan, P., Xu, Y., Jin, R.: Self-supervised video representation learning by context and motion decoupling. In: CVPR. pp. 13886–13895 (2021)
10. Huang, Z., Zhang, S., Jiang, J., Tang, M., Jin, R., Ang, M.H.: Self-supervised motion learning from static images. In: CVPR. pp. 1276–1285 (2021)
11. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: ICCV. pp. 2556–2563. IEEE (2011)

12. Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W.: Videomoco: Contrastive video representation learning with temporally adversarial examples. In: CVPR. pp. 11205–11214 (2021)
13. Patrick, M., Asano, Y.M., Huang, B., Misra, I., Metze, F., Henriques, J., Vedaldi, A.: Space-time crop & attend: Improving cross-modal video representation learning. In: ICCV (2021)
14. Rai, N., Adeli, E., Lee, K.H., Gaidon, A., Niebles, J.C.: Cocon: Cooperative-contrastive learning. In: CVPR. pp. 3384–3393 (2021)
15. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
16. Wang, J., Jiao, J., Liu, Y.H.: Self-supervised video representation learning by pace prediction. In: ECCV. pp. 504–521. Springer (2020)
17. Xiao, F., Tighe, J., Modolo, D.: Modist: Motion distillation for self-supervised video representation learning. arXiv preprint arXiv:2106.09703 (2021)
18. Yao, Y., Liu, C., Luo, D., Zhou, Y., Ye, Q.: Video playback rate perception for self-supervised spatio-temporal representation learning. In: CVPR. pp. 6548–6557 (2020)