Supplementary Material for Motion Sensitive Contrastive Learning for Self-supervised Video Representation

Jingcheng Ni[1,2], Nan Zhou[1,2], Jie Qin[3], Qian Wu[4], Junqi Liu[4], Boxun Li[4], and Di Huang[1,2]

[1] SKLSDE, Beihang University, Beijing, China
[2] SCSE, Beihang University, Beijing, China
[3] CCST, NUAA, Nanjing, China
[4] MEGVII Technology

1 Results on larger backbones

To evaluate the ability of modeling dynamics, we conduct experiments on the SSv2 dataset with 3D the ResNet-50 [2] (with less temporal down-sampling) backbone. As Table 1 shows, our method outperforms the previous state-of-the-art methods with the top-1 accuracy of 58.7%. In contrast to the K400 dataset, videos in SSv2 share similar appearances and backgrounds, and temporal information among different frames plays a more important role in the action classification task. The results on this dataset show that our method can indeed enhance temporal feature learning, which confirms the purpose of MSCL.

2 Analytical experiments

To further show the effectiveness of the proposed solutions, we conduct analytical experiments on the SSv2 dataset. We set the baseline by removing Local Motion Contrastive Learning (LMCL) and Motion Differential Sampling (MDS) from the entire method.

In the top 3 rows of Table 2, we give some classes that are most confusing to the baseline on which our method performs better. Obviously, motion

Table 1: Action classification results with R50 backbone and † indicates the model is trained with 3 views.

Method	Date	Dataset	Sizes	Epochs	SSv2
MoDist [5]	2021	K400	16×256	800	57.4
ρMoCo [1]†	2021	K400	8×256	200	54.4
CORP$_f$ [3]	2021	K400	16×256	200	41.1
CORP$_m$ [3]	2021	K400	16×256	200	48.8
Ours	-	K400	16×256	200	58.7

Table 2: Pairs of actions most confusing to the baseline (top-3 rows) and MSCL (bottom-3 rows), along with the number of videos that each model confuses. We use the relative proportion to rank confusing pairs.

Confusing pair	Confused videos Baseline	MSCL
Pouring something into something until it overflows Pouring something into something	26	**16**
Showing something on top of something Showing something behind something	50	**37**
Putting something and something on the table Putting something, something and something on the table	31	**18**
Rolling something on a flat surface Letting something roll along a flat surface	**51**	66
Moving something across a surface until it falls down Pushing something so that it falls off the table	**39**	45
Pretending to poke something Poking something so lightly that it (almost) doesn't move	**31**	37

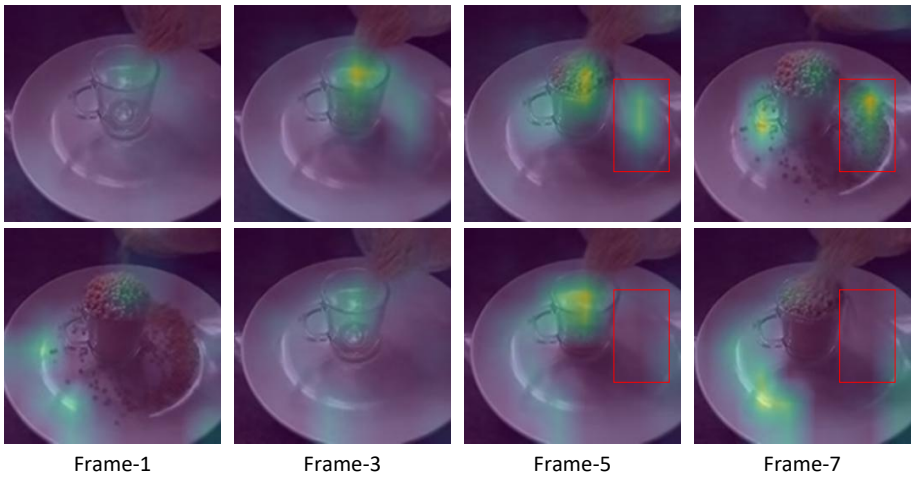

Fig. 1: Visualization of examples with label **"Pouring something into something until it overflows"**. The top row shows GradCAM of MSCL and the bottom row shows GradCAM of the baseline.

sensitive representation makes the model benefit from better temporal information. For example, separating action pairs like "Showing something on top of something" and "Showing something behind something", "Putting something

and something on the table" and "Putting something, something and something on the table" shows the improvement of distinguishing motion trajectories with fine-grained differences. The less confusing videos belonging to the pair: "Pouring something into something until it overflows" and "Pouring something into something" indicate that our method recognizes the final state of the action more accurately, which can be attributed to better local representation learned from LMCL. Moreover, in the bottom 3 rows of Table 2, we list the most confusing pairs of our method. We observe that these action pairs either have little motion information (*e.g.* "Pretending to poke something" and "Poking something so lightly that it (almost) doesn't move") or require appearance information for distinguishing (*e.g.* "Moving something across a surface until it falls down" and "Pushing something so that it falls off the table"). As better local temporal representation can take effect on the SSv2 dataset (*e.g.* higher top-1 accuracy), the appearance information is relatively ignored compared to the baseline.

We visualize the saliency map generated by GradCAM [4] to further show that our method can handle the confusing pairs mentioned above. In Fig. 1, we visualize the example with action "Pouring something into something until it overflows". Compared to the baseline, our method focuses on the motion region (marked by red bounding-boxes) corresponding to "overflow". Moreover, we find an interesting phenomenon in Fig. 2, where motion information in the 3-rd frame and 5-th frame can help recognizing action "Putting something behind something". Obviously, the baseline only focuses on one frame while MSCL takes advantage of both frames. This observation becomes the evidence that local temporal information is better explored by our method.

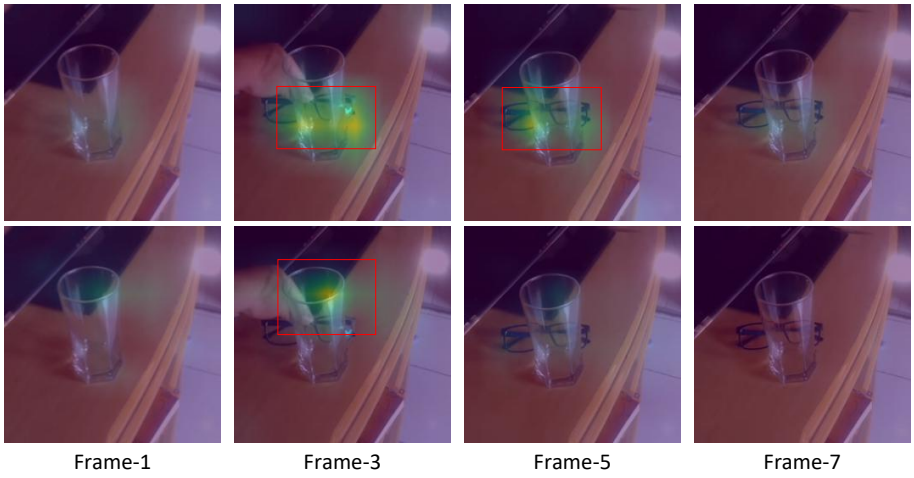

Fig. 2: Visualization of examples with label **"Putting something behind something"**. The top row shows GradCAM of MSCL and the bottom row shows GradCAM of the baseline.

References

1. Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K.: A large-scale study on unsupervised spatiotemporal representation learning. In: CVPR. pp. 3299–3309 (2021)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
3. Hu, K., Shao, J., Liu, Y., Raj, B., Savvides, M., Shen, Z.: Contrast and order representations for video self-supervised learning. In: ICCV. pp. 7939–7949 (2021)
4. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: CVPR. pp. 618–626 (2017)
5. Xiao, F., Tighe, J., Modolo, D.: Modist: Motion distillation for self-supervised video representation learning. arXiv preprint arXiv:2106.09703 (2021)