Supplementary Material: CMD: Self-supervised 3D Action Representation Learning with Cross-modal Mutual Distillation

Yunyao Mao[1], Wengang Zhou[1,2,*], Zhenbo Lu[2], Jiajun Deng[1], and Houqiang Li[1,2,*]

[1] CAS Key Laboratory of Technology in GIPAS, EEIS Department, University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
myy2016@mail.ustc.edu.cn, zhwg@ustc.edu.cn, luzhenbo@iai.ustc.edu.cn, dengjj@ustc.edu.cn, lihq@ustc.edu.cn

A Additional Ablation studies

In this section, we provide more ablative experiments on the NTU-60 dataset according to the cross-subject protocol.

Temperatures in CMD: Table 1 shows the performance of the learned representation for different values of temperature τ_t and τ_s. We can find that **i)** our CMD exhibits the optimal performance when using temperature $\tau_s = 0.1$ for the student and a smaller $\tau_t = 0.05$ for the teacher. **ii)** the performance does not show significant changes when τ_t varies between small values (from 0.01 to 0.05). **iii)** the learned representation gets worse as τ_s increases from 0.1 to 1.0.

Table 1. Ablative experiments of the temperatures in CMD. The performance is evaluated on the NTU-60 dataset according to the cross-subject protocol.

τ_t	0.05	0.07	0.10	0.01			0.02			0.05		
τ_s	0.05	0.07	0.10	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0
Accuracy (%)	77.8	78.0	77.5	79.7	78.2	76.8	79.7	77.8	76.5	**79.8**	78.0	76.8

One-stage vs. two-stage: We tried one-stage and two-stage training for both CrosSCLR-B [1] and our CMD. As shown in Table 2, CrosSCLR-B suffers severe performance drop with one-stage training, while CMD does not show significant performance gap between one-stage and two-stage training.

Mutual contrastive loss vs. mutual distillation: We conduct bidirectional mutual contrastive learning (denote as MCL) for all three modalities and compare its performance with our CMD. As shown in Table 3, CMD outperforms MCL under both linear and KNN protocols, showing the superiority of mutual knowledge distillation.

* Corresponding authors: Wengang Zhou and Houqiang Li

Table 2. One-stage training vs. two-stage training.

Dataset	**CrosSCLR-B**		**CMD (Ours)**	
	one-stage	two-stage	one-stage	two-stage
NTU-60 x-sub	57.6	77.3	**79.4**	79.2

Table 3. Mutual contrastive loss vs. mutual distillation.

Dataset	**Linear Evaluation**			**KNN Evaluation**		
	Base	MCL	CMD	Base	MCL	CMD
NTU-60 x-sub	76.1	77.3	**79.4**	63.4	64.3	**70.6**

B Limitations and future work

(1) The complexity of cross-modal mutual distillation is proportional to the square of the number of modalities. How to design a distillation framework with linear computational complexity remains a promising direction; (2) CMD benefits from the complementarity between skeleton modalities. When applied to other CV domains, multiple good but different modalities (e.g. RGB, Depth, and Events) that complement each other are required, which may be somewhat hard to collect.

References

1. Li, L., Wang, M., Ni, B., Wang, H., Yang, J., Zhang, W.: 3d human action representation learning via cross-view consistency pursuit. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4741–4750 (2021) 1