On Compositions of Transformations in Contrastive Self-Supervised Learning

UNIVERSITY OF OXFORD

FACEBOOK

2021 ICCV OCTOBER 11-17 VIRTUAL

Mandela Patrick*, Yuki M. Asano*,
Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, Andrea Vedaldi

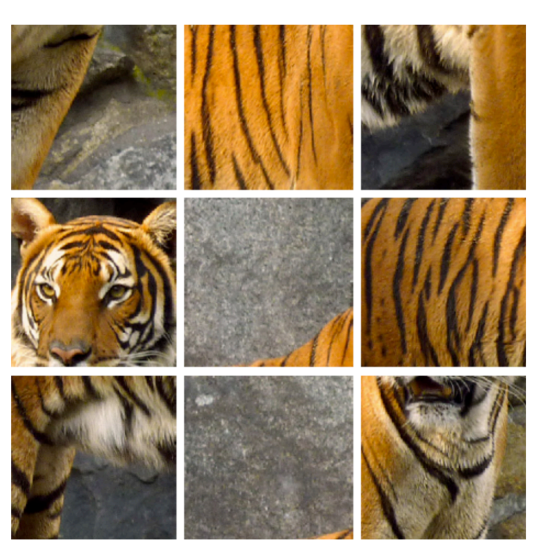
Abstract.

In the image domain, excellent representations can be learned by inducing invariance to content-preserving trans-formations via noise contrastive learning. In this paper, we generalize contrastive learning to a wider set of transformations, and their compositions, for which either invariance or distinctiveness is sought. We show that it is not immediately obvious how existing methods such as SimCLR can be extended to do so. Instead, we introduce a number of formal requirements that all contrastive formulations must satisfy, and propose a practical construction which satisfies these requirements. In order to maximise the reach of this analysis, we express all components of noise contrastive formulations as the choice of certain generalized transformations of the data (GDTs), including data sampling. We then consider videos as an example of data in which a large variety of transformations are applicable, accounting for the extra modalities – for which we analyze audio and text– and the dimension of time. We find that being invariant to certain transformations and distinctive to others is critical to learning effective video representations, improving the state-of-the-art for multiple benchmarks by a large mar-gin, and even surpassing supervised pretraining.

Code and pretrained models at: https://github.com/facebookresearch/GDT

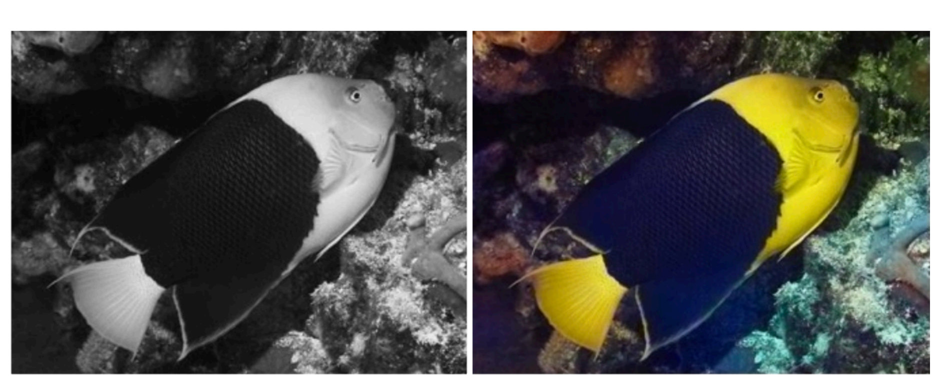
Self-supervision = learning *invariance* to some transformations, *variance* to others.

e.g.

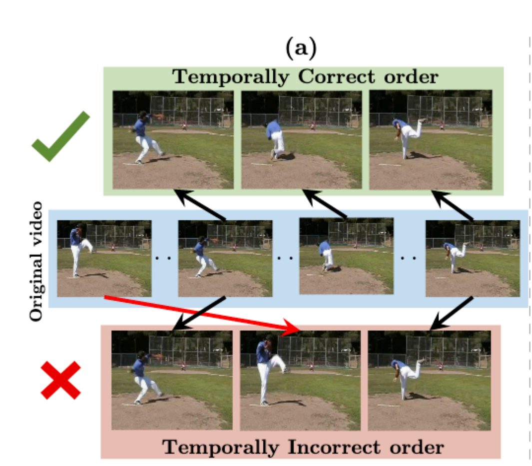


Jigsaw

RotNet

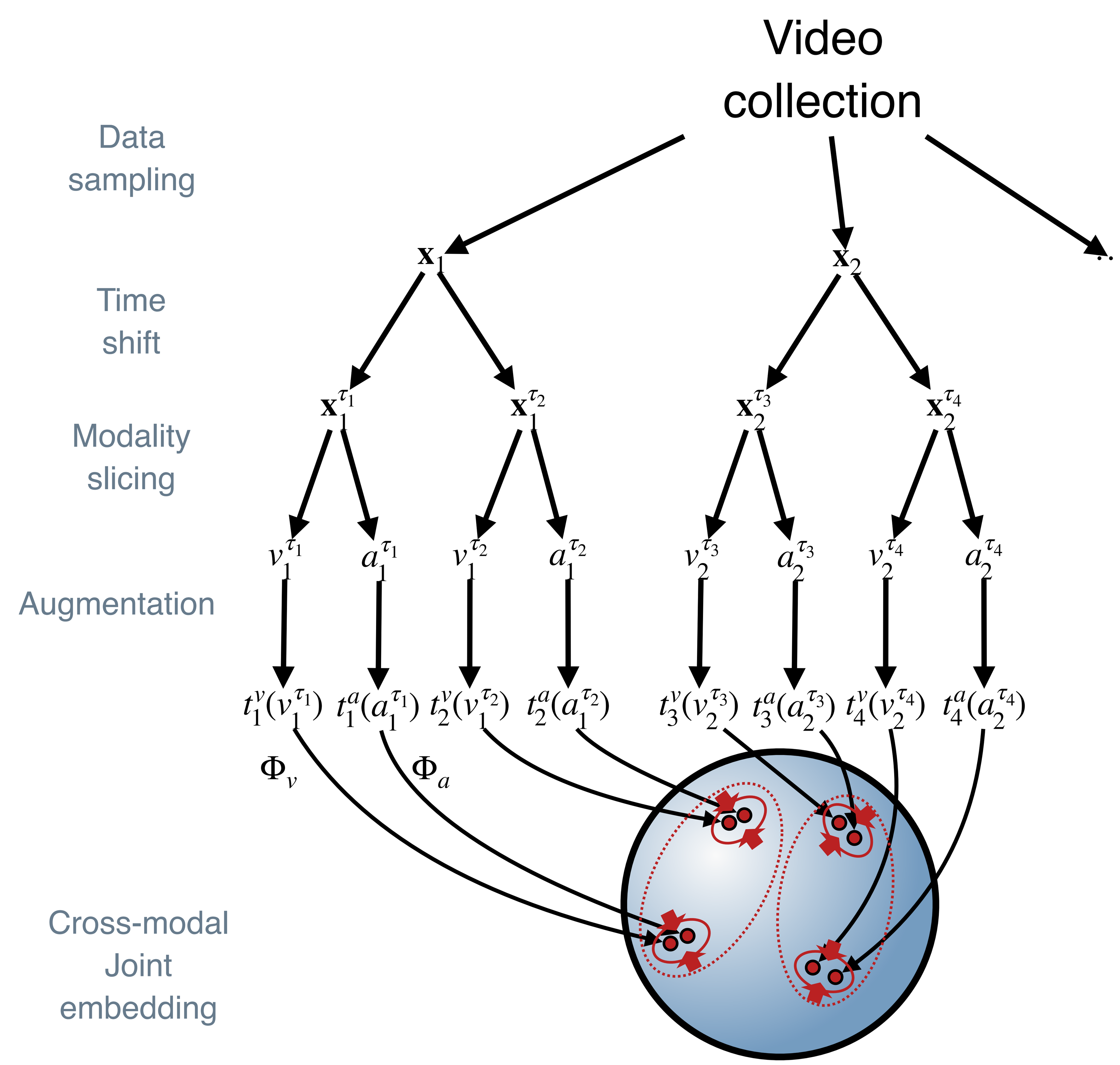


Colorization

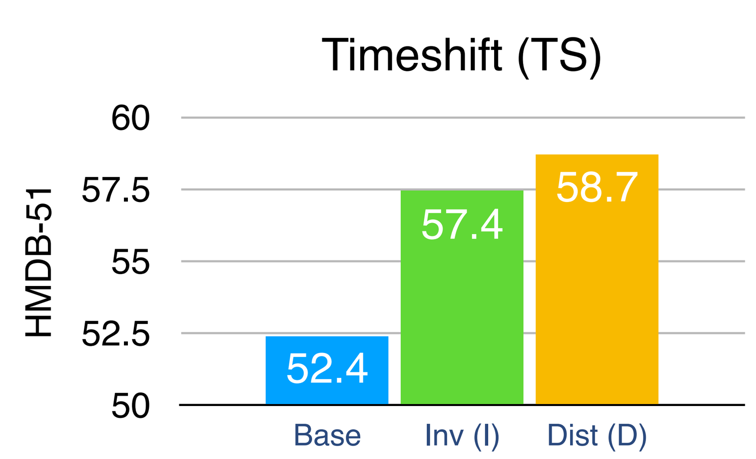


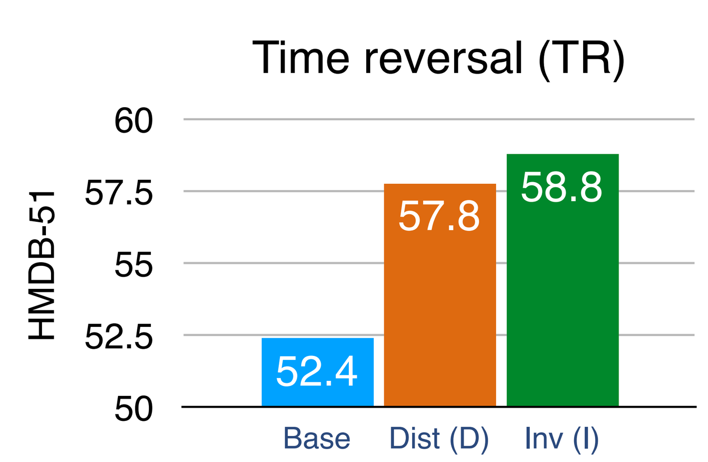
Video frames shuffling

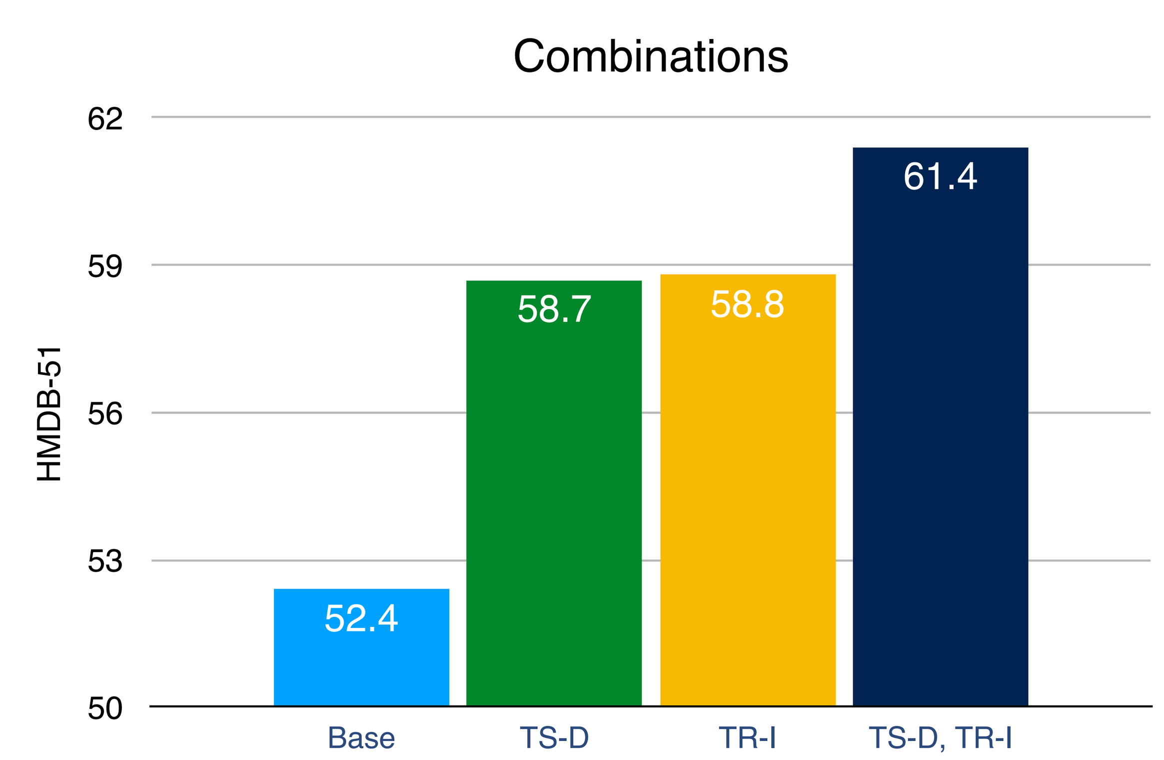
Hierarchical sampling



Test different & novel learning hypotheses

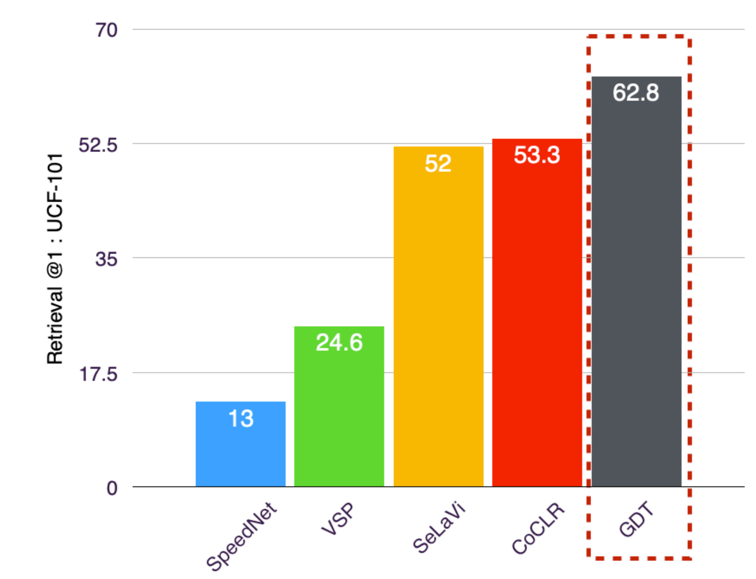


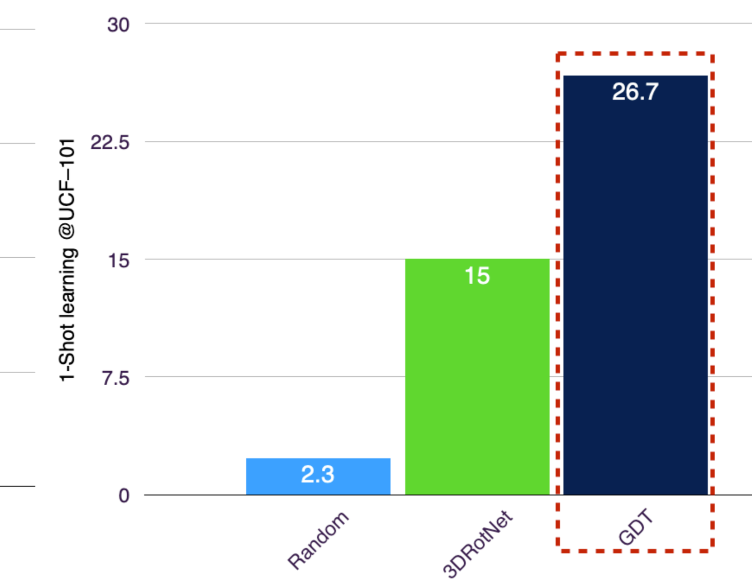




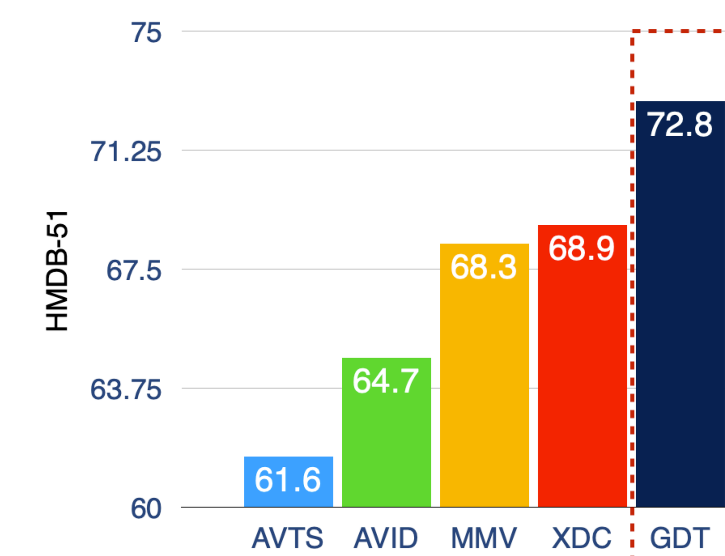
State-of-the-art representation learning results

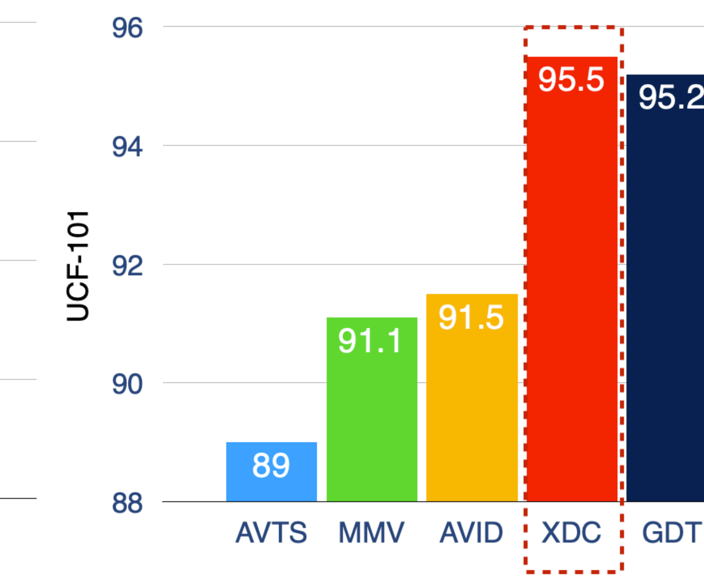
SOTA video action retrieval and few-shot learning results





SOTA finetuning video-action recognition results





Code and pretrained models

https://github.com/facebookresearch/GDT

References

Noroozi et al. Unsupervised earning of visual representations by solving jigsaw puzzles. ECCV 2016.
Gidaris et al. Unsupervised representation learning by predicting image rotations. ICLR 2018.
Zhang, et al. Colorful image colorization. ECCV 2016.
Misra et al. Shuffle and learn: unsupervised learning using temporal order verification. ECCV 2016.
Miech et al. End-to-end learning of visual representations from uncurated instructional videos. CVPR 2020,
Korbar et al. Cooperative learning of audio and video models from self-supervised synchronization. NeurIPS 2018
Alwassel et al. Self-supervised learning by cross-modal audio-video clustering. NeurIPS 2020
Jing et al. Self-supervised spatiotemporal feature learning by video geometric transformations. arXiv.
Chu et al. Self-supervised spatiotemporal learning via video clip order prediction. CVPR 2019.
Benaim et al. Speednet: Learning the speediness in videos. CVPR 2020
Cho et al. Self-supervised spatio-temporal representation learning using variable playback speed prediction. IEEE Access.
Luo et al. Video cloze procedure for self-supervised spatio-temporal learning. AAAI 2020