Supplementary Material for:
Time-Equivariant Contrastive Video Representation Learning

Simon Jenni Hailin Jin
Adobe Research
{jenni,hljin}@adobe.com

1. Additional Implementation Details

We provide some additional implementation details regarding network architectures, data augmentations, and evaluation protocol.

Network Architectures. We noticed some inconsistencies regarding the definition of the 3D-ResNet architecture used in prior works. Our 3D-Resnet architecture is identical to the one used in [10, 11, 13] with the first two residual blocks consisting only of 2D convolutions and the final two blocks consisting of 3D convolutions. The R(2+1)D and S3D-G architectures are identical to the original works [26, 29].

Data Augmentation Details. Our temporal augmentations consist of 1) choosing a temporal subsampling factor corresponding to $1\times$, $2\times$, $4\times$ or $8\times$ playback, 2) randomly choosing forward or backward playback with equal probability 3) randomly sampling k consecutive frames satisfying the constraints of steps 1 and 2.

We use a standard augmentation pipeline for the spatial and color jittering as found in contrastive learning methods [5]. Concretely, we sample crops with an area covering δ-times the original area, with δ chosen randomly in the range $[0.2, 1.0]$. Similarly, the crop aspect ratio is chosen randomly from the range $[3/4, 4/3]$. Finally, we apply random horizontal flipping with a probability of 0.5.

Color jittering consists of random modifications of brightness, saturation, hue, and contrast. These random modifications are performed in random order to add more variability. Finally, we randomly convert videos to greyscale with a probability of 0.2. All the spatial and color jitterings are applied consistently to all frames of the video. We do not use random Gaussian blur in our experiments.

Evaluation Details. As mentioned in Section 3.4, we performed a multi-crop evaluation with a combination of temporal and spatial crops. We followed the same approach as [12] and uniformly sampled ten temporal crops and additionally extracted ten spatial crops each (*i.e.*, center crops + four corner crops and each also with horizontal flipping). Spatial crops were extracted at a resolution of 176×176 for R(2+1)D and 192×192 for R3D-18 and S3D-G (adjusting for the different pre-training resolutions). Since averaging over more crops can impact the final performance and not all the prior works follow this protocol, we also include numbers obtained with only a single spatial crop in the extended comparison Tables 1 & 2.

2. Additional Results

Qualitative Nearest-Neighbor Results. We illustrate some qualitative nearest neighbor retrievals on UCF101 obtained with the R3D-18 network in Figure 1. The nearest neighbor computation is again based on cosine similarity on standardized feature vectors, as described in Section 3.4. The sensible retrievals reflect the excellent performance in the video retrieval evaluation (see Table 2).

Additional Comparisons to Prior Work. We report additional comparisons in transfer towards action recognition in Table 1. We include results obtained with an R(2+1)D using UCF101 pre-training in the top block, observing improvements over prior works in the same setting. Additionally, we report results obtained without using multiple spatial crops, allowing for a fairer comparison to prior works using only single spatial crops.

Finally, we report additional comparisons on the video retrieval task in Table 2. Besides the single crop evaluation, we also report performance obtained with Kinetics pre-training for completeness.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019. 3

[2] Yutong Bai, Haoqi Fan, Ishan Misra, Ganesh Venkatesh, Yongyi Lu, Yuyin Zhou, Qihang Yu, Vikas Chandra, and

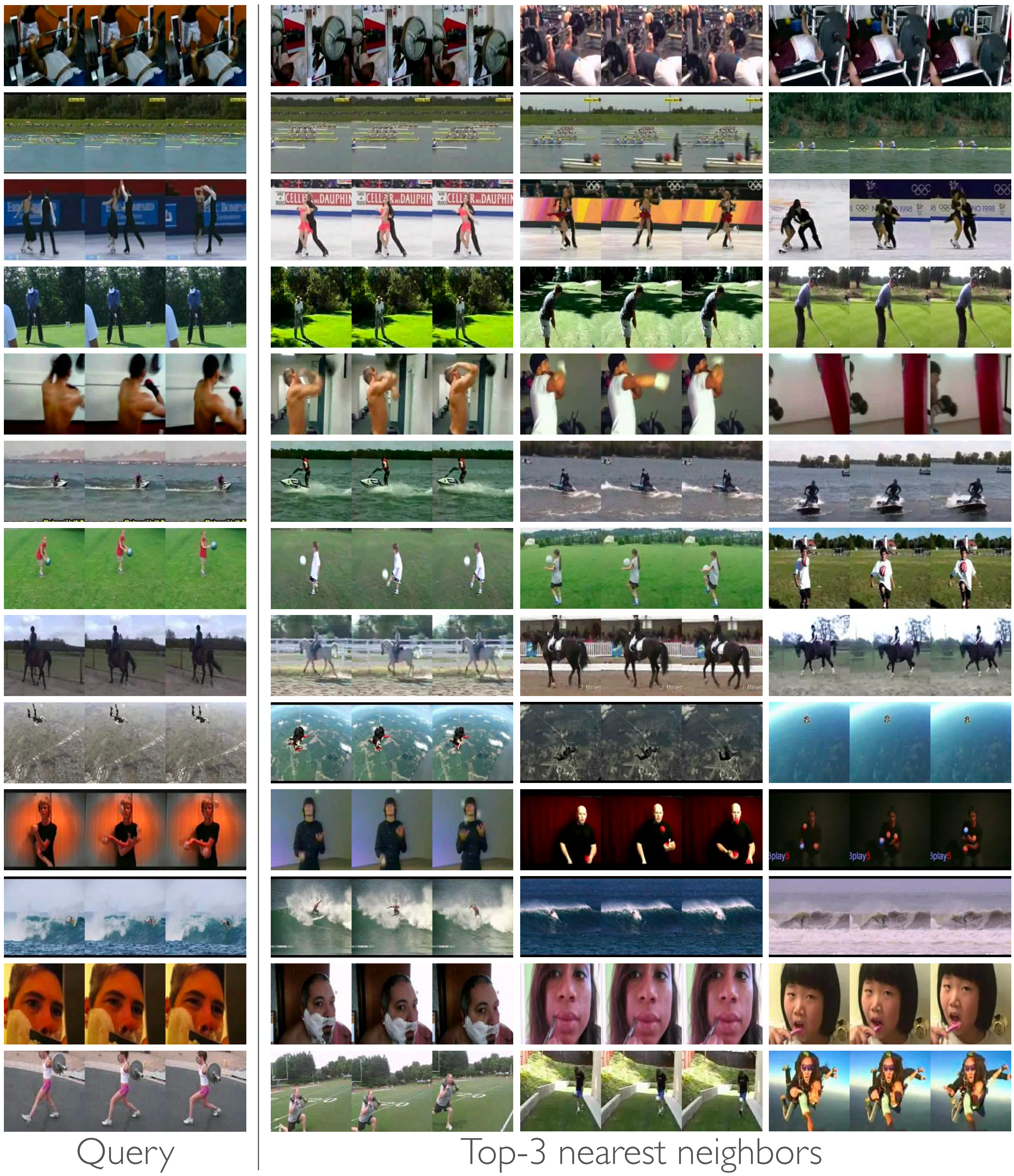

Figure 1. **Qualitative nearest-neighbor retrievals on UCF101.** We show three frames from a query video on the left (taken from UCF101 test split 1), followed by the three nearest neighbors from the training set (train split 1). Nearest neighbors are computed using cosine similarity in the feature space of a 3D-ResNet pre-trained using our proposed self-supervised learning task.

Table 1. **Extended comparison to prior work on self-supervised video representation learning.** We report action recognition accuracy after fine-tuning to UCF101 and HMDB51. We indicate the pre-training dataset, input resolution, number of input frames, network architecture, and pre-training data modality (V=RGB, F=optical-flow, A=audio, T=text). In the upper block we compare to other methods with an R(2+1)D network when pre-training on UCF101. We also report numbers obtained with single spatial crop evaluation since not all methods perform inference using multiple spatial crops.

Method	**Dataset**	**Res.**	**Frames**	**Network**	**Mod.**	**UCF101**	**HMDB51**
VCP [18]	UCF101	112	16	R(2+1)D	V	66.3	32.2
PRP [32]	UCF101	112	16	R(2+1)D	V	72.1	35.0
VCOP [30]	UCF101	112	16	R(2+1)D	V	72.4	30.9
STS [27]	UCF101	112	16	R(2+1)D	V	73.6	34.1
Var. PSP [6]	UCF101	112	16	R(2+1)D	V	74.8	36.8
Pace Pred. [28]	UCF101	112	16	R(2+1)D	V	75.9	35.9
Temp.-Trans. [13]	UCF101	112	16	R(2+1)D	V	81.6	46.4
TCRL [7]	UCF101	112	16	R(2+1)D	V	82.8	53.6
Ours (1-crop)	UCF101	112	16	R(2+1)D	V	85.2	56.9
Ours (10-crop)	UCF101	112	16	R(2+1)D	V	85.8	59.3
3D ST-puzzle [15]	Kinetics-400	224	16	R3D-18	V	65.8	33.7
3D RotNet [14]	Kinetics-400	112	16	R3D-18	V	66.0	37.1
STS [27]	Kinetics-400	112	16	R3D-18	V	68.1	34.4
Temp.-Trans. [13]	Kinetics-400	112	16	R3D-18	V	79.3	49.8
DPC [10]	Kinetics-400	224	40	R3D-34	V	75.7	35.7
MemDPC [11]	Kinetics-400	224	40	R3D-34	V	78.1	41.2
Pace Pred. [28]	Kinetics-400	112	16	R(2+1)D	V	77.1	36.6
VideoMoCo [21]	Kinetics-400	112	16	R(2+1)D	V	78.7	49.2
VideoDIM [8]	Kinetics-400	128	32	R(2+1)D	V	79.7	49.2
TCRL [7]	Kinetics-400	112	16	R(2+1)D	V	84.3	54.2
CBT [24]	Kinetics-600	112	16	S3D	V	79.5	44.6
SpeedNet [3]	Kinetics-400	224	64	S3D-G	V	81.1	48.8
VTHCL [31]	Kinetics-400	224	8	R50	V	82.1	49.2
TaCo [2]	Kinetics-400	224	16	R50	V	85.1	51.6
CVRL [23]	Kinetics-400	224	32	R3D-50	V	92.1	65.4
DynamoNet [9]	Youtube8M	112	32	STCNet	V	88.1	59.9
STS [27]	Kinetics-400	224	64	S3D-G	V+F	89.0	62.0
CoCRL [12]	Kinetics-400	128	32	S3D	V+F	87.9	54.6
AVTS [16]	Kinetics-400	224	25	MC3	V+A	85.8	56.9
XDC [1]	Kinetics-400	224	8	R(2+1)D	V+A	84.2	47.1
GDT [22]	Kinetics-400	112	32	R(2+1)D	V+A	89.3	60.0
MIL-NCE [19]	HowTo100M	224	32	S3D	V+T	91.3	61.0
Ours (1-crop)	Kinetics-400	128	16	R3D-18	V	85.5	60.9
Ours (1-crop)	Kinetics-400	112	16	R(2+1)D	V	87.1	59.8
Ours (1-crop)	Kinetics-400	128	32	S3D-G	V	86.3	58.6
Ours (10-crop)	Kinetics-400	128	16	R3D-18	V	87.1	63.6
Ours (10-crop)	Kinetics-400	112	16	R(2+1)D	V	88.2	62.2
Ours (10-crop)	Kinetics-400	128	32	S3D-G	V	86.9	63.5

Alan Yuille. Can temporal information help with contrastive self-supervised learning? *arXiv preprint arXiv:2011.13046*, 2020. 3

[3] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020. 3, 4

[4] Uta Büchler, Biagio Brattoli, and Björn Ommer. Improving spatiotemporal self-supervision by deep reinforcement learning. *arXiv preprint arXiv:1807.11293*, 2018. 4

Table 2. **Extended comparison on the video retrieval tasks on UCF101 and HMDB51.** We report recall at k (R@k) for k-NN based video retrieval. Query videos are taken from test split 1 and retrievals computed on train split 1 of UCF101 and HMDB, respectively. * indicates Kinetics pre-training. We also report results when using a single spatial crop to extract feature vectors (instead of averaging over ten spatial crops).

		UCF101				HMDB51			
Method	**Network**	**R@1**	**R@5**	**R@10**	**R@20**	**R@1**	**R@5**	**R@10**	**R@20**
Jigsaw [20]	AlexNet	19.7	28.5	33.5	40.0	-	-	-	-
OPN [17]	AlexNet	19.9	28.7	34.0	40.6	-	-	-	-
Büchler *et al.* [4]	AlexNet	25.7	36.2	42.2	49.2	-	-	-	-
STS [27]	C3D	30.1	49.6	58.8	67.6	13.9	33.3	44.7	59.5
Pace Pred. [28]	C3D	31.9	49.7	59.2	68.9	12.5	32.2	45.4	61.0
PRP [32]	R3D-18	22.8	38.5	46.7	55.2	-	-	-	-
VCOP [30]	R3D-18	14.1	30.3	40.4	51.1	7.6	22.9	34.4	48.0
VCP [18]	R3D-18	18.6	33.6	42.5	53.5	7.6	24.4	36.6	53.6
Var. PSP [6]	R3D-18	24.6	41.9	51.3	62.7	10.3	26.6	38.8	51.6
PCL [25]	R3D-18	40.5	59.4	68.9	77.4	16.8	38.4	53.4	68.9
MemDPC [11]	R3D-18	20.2	40.4	52.4	64.7	7.7	25.7	40.6	57.7
Temp.-Trans. [13]*	R3D-18	26.1	48.5	59.1	69.6	-	-	-	-
SpeedNet [3]*	S3D-G	13.0	28.1	37.5	49.5	-	-	-	-
CoCRL [12]	S3D	53.3	69.4	76.6	82.0	23.2	43.2	53.5	65.5
TCRL [7]	R(2+1)D	56.9	72.2	79.0	84.6	24.1	45.8	58.3	75.3
GDT [22]*	R(2+1)D	57.4	73.4	80.8	88.1	25.4	51.4	63.9	75.0
Ours (1-crop)	R3D-18	62.5	78.4	84.1	88.8	32.0	60.8	72.2	81.7
Ours (1-crop)	R(2+1)D	64.6	80.8	85.8	90.5	29.7	53.7	66.9	77.8
Ours (10-crop)	R3D-18	63.6	79.0	84.8	89.9	32.2	60.3	71.6	81.5
Ours (10-crop)	R(2+1)D	64.3	80.9	86.4	90.6	29.5	55.8	68.0	78.2
Ours (1-crop)*	R3D-18	66.9	83.1	88.8	93.3	36.4	64.1	74.1	83.8
Ours (1-crop)*	R(2+1)D	64.2	81.0	87.6	92.4	33.2	59.7	72.4	82.9
Ours (10-crop)*	R3D-18	67.8	83.7	88.9	93.7	38.0	65.2	75.9	83.2
Ours (10-crop)*	R(2+1)D	64.2	81.1	87.4	92.6	33.1	60.8	73.1	84.1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[6] Hyeon Cho, Taehoon Kim, Hyung Jin Chang, and Wonjun Hwang. Self-supervised spatio-temporal representation learning using variable playback speed prediction. *arXiv preprint arXiv:2003.02692*, 2020. 3, 4

[7] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *arXiv preprint arXiv:2101.07974*, 2021. 3, 4

[8] R Devon et al. Representation learning with video deep infomax. *arXiv preprint arXiv:2007.13278*, 2020. 3

[9] Ali Diba, Vivek Sharma, Luc Van Gool, and Rainer Stiefelhagen. Dynamonet: Dynamic action and motion network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6192–6201, 2019. 3

[10] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 3

[11] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020. 1, 3, 4

[12] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020. 1, 3, 4

[13] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. *arXiv preprint arXiv:2007.10730*, 2020. 1, 3, 4

[14] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018. 3

[15] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019. 3

[16] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018. 3

[17] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sort-

ing sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 4

[18] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11701–11708, 2020. 3, 4

[19] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 3

[20] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 4

[21] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. *arXiv preprint arXiv:2103.05905*, 2021. 3

[22] Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multimodal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020. 3, 4

[23] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *arXiv preprint arXiv:2008.03800*, 2020. 3

[24] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019. 3

[25] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Self-supervised video representation using pretext-contrastive learning. *arXiv preprint arXiv:2010.15464*, 2020. 4

[26] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1

[27] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Wei Liu, and Yun-hui Liu. Self-supervised video representation learning by uncovering spatio-temporal statistics. *arXiv preprint arXiv:2008.13426*, 2020. 3, 4

[28] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European Conference on Computer Vision*, pages 504–521. Springer, 2020. 3, 4

[29] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018. 1

[30] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 3, 4

[31] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. Video representation learning with visual tempo consistency. *arXiv preprint arXiv:2006.15489*, 2020. 3

[32] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6548–6557, 2020. 3, 4