Supplementary material for: Motion-Augmented Self-Training for Video Recognition at Smaller Scale

Kirill Gavrilyuk[1] Mihir Jain[2] Ilia Karmanov[2] Cees G. M. Snoek[1]

[1]University of Amsterdam [2]Qualcomm AI Research*

{kgavrilyuk,cgmsnoek}@uva.nl {mijain,ikarmano}@qti.qualcomm.com

In this supplementary material, we first report qualitative results for video clip retrieval. Then we provide more analysis on semi-supervised action recognition. Finally, we make a larger comparison with self-supervised methods that use other than Kinetics dataset[1] for training.

1. Qualitative results for video retrieval

To further investigate the quality of the learned representation, we illustrate a few success and failure examples in Figure 1. Despite that some of the retrieved videos are from different action classes than the query video, the learned representation successfully captures similar motion patterns and not the appearance context. For example, on the first row the model captures hand motion, on the second row the model captures the dominant human poses.

2. Semi-supervised action recognition

We further analyze our method for semi-supervised video recognition with ablation on the size of source dataset (D_{source}) for self-training. The results are shown in Table 1. In the main paper, following competing methods, we use a labelled set (20% or 50% of training set) as D_{MPLG} and the full training set as D_{source}. Here, we experiment by removing labelled samples from the training set to have a trimmed D_{source} = training set$-D_{MPLG}$, with the remaining 80% or 50% samples of the training set. With this, the performance drops for both labeled subsets but the larger drop of 2.4% for a 50% labelled subset is due to the drastic decrease in the size of D_{source}. While the 20% labelled subset loses only 0.1% as the data for self-training is only slightly reduced. This further shows the importance of our self-training.

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

[1]Datasets used in this paper were downloaded and experimented on by primary author

	UCF101, split 1		
	20% labelled	50% labelled	D_{source} / Training dataset
Jing *et al.* [15]‡	48.7	54.3	training set
Rizve *et al.* [29]	39.4	50.2	training set
MotionFit (*ours*)	57.7	59.0	training set
MotionFit (*ours*)	57.6	56.6	training set $-D_{MPLG}$

‡ Use extra labels to pre-train a 2D CNN.

Table 1: **Comparison with semi-supervision** on video recognition at smaller scale. We report top-1 accuracy of models fine-tuned on 20% (or 50%) of UCF101 training data, which is also our D_{MPLG} (same as D_{target}).

3. Comparison with self-supervised methods for action recognition

In Table 2 we list more results of self-supervised methods for action recognition that also utilize datasets other than Kinetics for training. Our method still outperforms other visual-only methods, including those that use a larger dataset for pre-training [8].

4. Comparison with self-supervised methods for clip retrieval

In Table 3 we list more results of self-supervised methods for clip retrieval that also utilize other than Kinetics dataset for training. Our method still outperforms other visual-only methods.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, A. Natsev, G. Toderici, B. Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *ArXiv*, abs/1609.08675, 2016. 2

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 2

Figure 1: **Success and failure examples** for video retrieval. The left examples show top-retrieved videos belonging to the same action class as the query video. On the right, all three retrieved videos are from a different class. While our model may retrieve videos from different action classes, it still captures distinctive motion patterns like hand motion and human poses.

	Dataset	Backbone	Frames	Resolution	Modality	UCF101	HMDB51
Sun *et al.* [30]	K400	S3D	16	112	V + T	79.5	44.6
Owens *et al.* [25]	K400	R3D-18	64	224	V + A	82.1	-
Asano *et al.* [3]	K400	R(2+1)D-18	30	112	V + A	83.1	47.1
Asano *et al.* [3]	VGG-Sound [6]	R(2+1)D-18	30	112	V + A	87.7	53.1
Korbar *et al.* [19]	K400	MC3-18	25	224	V + A	85.8	56.9
Korbar *et al.* [19]	Audioset [10]	MC3-18	25	224	V + A	89.0	61.6
Alwassel *et al.* [2]	K400	R(2+1)D-18	32	224	V + A	86.8	52.6
Alwassel *et al.* [2]	Audioset [10]	R(2+1)D-18	32	224	V + A	93.0	63.7
Alwassel *et al.* [2]	IG-Kinetics [11]	R(2+1)D-18	32	224	V + A	95.5	68.9
Xiao *et al.* [31]	K400	SlowFast	64	224	V + A	87.0	54.6
Morgado *et al.* [23]	K400	R(2+1)D-18	32	224	V + A	87.5	60.8
Morgado *et al.* [23]	Audioset [10]	R(2+1)D-18	32	224	V + A	91.5	64.7
Patrick *et al.* [26]	K400	R(2+1)D-18	32	224	V + A	89.3	60.0
Patrick *et al.* [26]	VGG-Sound [6]	R(2+1)D-18	32	224	V + A	89.4	62.1
Patrick *et al.* [26]	Audioset [10]	R(2+1)D-18	32	224	V + A	92.5	66.1
Patrick *et al.* [26]	IG-Kinetics [11]	R(2+1)D-18	32	224	V + A	95.2	72.8
Miech *et al.* [21]	HTM [22]	S3D	32	224	V + T	91.3	61.0
Piergiovanni *et al.* [27]	Youtube8M [1]	S3D	32	224	V + T	93.8	67.4
ElNouby *et al.* [9]	UCF101	R3D-18	16	112	V	64.4	-
Kim *et al.* [17]	K400	R3D-18	16	112	V	65.8	33.7
Kong *et al.* [18]	K400	R3D-18	8	112	V	69.4	37.8
Luo *et al.* [20]	UCF101	R(2+1)D-18	16	112	V	66.3	32.2
Yao *et al.* [33]	UCF101	R(2+1)D-18	16	112	V	72.1	35.0
Xu *et al.* [32]	UCF101	R(2+1)D-18	16	112	V	72.4	30.9
Cho *et al.* [7]	UCF101	R(2+1)D-18	16	112	V	74.8	36.8
Han *et al.* [12]	K400	R-2D3D-34	25	224	V	75.7	35.7
Jing *et al.* [16]	K400	R3D-18	64	112	V	76.6	47.0
Zhuang *et al.* [34]	K400	SlowFast	16	112	V	77.0	46.5
Han *et al.* [13]	K400	R-2D3D-18	25	224	V	78.1	41.2
Benaim *et al.* [4]	K400	S3D-G	64	224	V	81.1	48.8
Han *et al.* [14]	UCF101	S3D	32	128	V	81.4	52.1
Han *et al.* [14]	K400	S3D	32	128	V	87.9	54.6
Diba *et al.* [8]	Youtube8M [1]	STCNet	32	112	V	88.1	59.9
Qian *et al.* [28]	K400	R3D-50	16	224	V	92.2	66.7
MotionFit (*ours*)	K400	R(2+1)D-18	32	112	V	88.9	61.4
MotionFit (*ours*)	K400	S3D-G	64	224	V	90.1	50.6

Table 2: **Comparison with self-supervised methods on video action recognition.** We report top-1 accuracy of fine-tuned models averaged over all 3 splits of UCF101 and HMDB51. Our approach is the best when only considering the visual modality (V).

[3] Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, 2020. 2, 3

[4] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. SpeedNet: Learning the speediness in videos. In *CVPR*, 2020. 2, 3

[5] U. Büchler, B. Brattoli, and B. Ommer. Improving spatiotem-

	Dataset	Backbone	Modality	UCF101			HMDB51		
				R@1	R@5	R@20	R@1	R@5	R@20
Asano *et al*. [3]	K400	R(2+1)D-18	V + A	52.0	68.6	84.5	24.8	47.6	75.5
Patrick *et al*. [26]	K400	R(2+1)D-18	V + A	57.4	73.4	88.1	25.4	51.4	75.0
Xu *et al*. [32]	UCF101	R(2+1)D-18	V	10.7	25.9	47.3	5.7	19.5	45.8
Benaim *et al*. [4]	K400	S3D-G	V	13.0	28.1	49.5	-	-	-
Noroozi *et al*. [24]	UCF101	AlexNet	V	19.7	28.5	40.0	-	-	-
Luo *et al*. [20]	UCF101	R(2+1)D-18	V	19.9	33.7	50.5	6.7	21.3	49.2
Han *et al*. [13]	UCF101	R(2+1)D-18	V	20.2	40.4	64.7	7.7	25.7	57.7
Yao *et al*. [33]	UCF101	R(2+1)D-18	V	20.3	34.0	51.7	8.2	25.3	51.0
Kong *et al*. [18]	K400	R3D-18	V	22.0	39.1	56.3	-	-	-
Cho *et al*. [7]	UCF101	R3D-18	V	24.6	41.9	62.7	10.3	26.6	54.6
Buchler *et al*. [5]	UCF101	CaffeNet	V	25.7	36.2	49.2	-	-	-
Han *et al*. [14]	UCF101	S3D	V	53.3	69.4	82.0	23.2	43.2	65.5
MotionFit (*ours*)	K400	S3D-G	V	31.6	51.7	70.3	-	-	-
MotionFit (*ours*)	K400	R(2+1)D-18	V	61.6	75.6	85.5	29.4	46.5	66.7

Table 3: **Comparison with self-supervised methods on video clip retrieval.** We report recall values R@n for n = 1, 5, 20 on UCF101 and HMDB51 split 1. Our approach is best when only considering the visual modality and on par with methods that use an additional audio modality during training.

poral self-supervision by deep reinforcement learning. In *ECCV*, 2018. 3

[6] Honglie Chen, Weidi Xie, A. Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 2

[7] Hyeon Cho, Tae-Hoon Kim, Hyung Jin Chang, and Wonjun Hwang. Self-supervised spatio-temporal representation learning using variable playback speed prediction. *arXiv preprint arXiv:2003.02692*, 2020. 2, 3

[8] Ali Diba, V. Sharma, L. Gool, and R. Stiefelhagen. Dynamonet: Dynamic action and motion network. In *ICCV*, 2019. 1, 2

[9] Alaaeldin El-Nouby, Shuangfei Zhai, Graham W. Taylor, and Joshua M. Susskind. Skip-Clip: Self-supervised spatiotemporal representation learning by future clip order ranking. *arXiv preprint arXiv:1910.12770*, 2019. 2

[10] J. Gemmeke, D. Ellis, Dylan Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 2

[11] Deepti Ghadiyaram, Matt Feiszli, Du Tran, Xueting Yan, Heng Wang, and D. Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, 2019. 2

[12] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *ICCVw*, 2019. 2

[13] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *ECCV*, 2020. 2, 3

[14] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *NeurIPS*, 2020. 2, 3

[15] Longlong Jing, Toufiq Parag, Zhe Wu, Yingli Tian, and Hongcheng Wang. VideoSSL: Semi-supervised learning for video classification. In *WACV*, 2021. 1

[16] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018. 2

[17] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, 2019. 2

[18] Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, and Tomokazu Murakami. Cycle-contrast for self-supervised video representation learning. In *NeurIPS*, 2020. 2, 3

[19] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018. 2

[20] Dezhao Luo, Chang Liu, Y. Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *AAAI*, 2020. 2, 3

[21] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, I. Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. 2020. 2

[22] Antoine Miech, D. Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, I. Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 2

[23] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. *arXiv preprint arXiv:/2004.12943*, 2020. 2

[24] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 3

[25] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 2

[26] Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020. 2, 3

[27] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving losses for unsupervised video representation learning. In *CVPR*, 2020. 2

[28] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, 2021. 2

[29] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2021. 1

[30] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 2

[31] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[32] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, 2019. 2, 3

[33] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *CVPR*, 2020. 2, 3

[34] Chengxu Zhuang, Alex Andonian, and D. Yamins. Unsupervised learning from video with deep neural embeddings. In *CVPR*, 2020. 2