Appendix

A. Generalization on Diving48

To further highlight the generalizability of our method to new domains and fine-grained actions, we finetune and evaluate with the challenging Diving48 dataset [14]. It contains 18K trimmed videos for 48 different diving sequences all of which take place in similar backgrounds and need to be distinguished by subtle differences such as the number of somersaults or the starting position. We use standard train/test split and report top-1 accuracy.

In Table 1, we show the performance of our model when pretrained on the full Kinetics-400 and on Mini-Kinetics (†). We compare these results to no pretraining, the temporal contrastive baseline pretrained on Kinetics-400, and supervised pretraining on Kinetics-400 with labels. Our method increases the performance over training from scratch by 7.9% and the temporal contrastive baseline by 6.6%. Our method even outperforms the supervised pretraining baseline by 4.5%. This suggests that by contrasting tubelets with different motions, our method is able to learn better video representations for fine-grained actions than supervised pretraining on Kinetics. When pretraining on Mini-Kinetics (3x smaller than Kinetics-400) the performance of our model does not decrease, again demonstrating the data efficiency of our approach.

Pretraining	Top-1
Supervised [24]	84.5
None	81.1
Temporal Contrast Baseline	82.4
This paper †	**89.4**
This paper	**89.0**

Table 1: **Generalization on Diving48 [14].** Comparison with temporal contrastive pretraining and supervised pretraining on Diving48. All models use R(2+1)D-18. † indicates pretraining on Mini-Kinetics, otherwise all pretraining was done on Kinetics-400.

B. Evaluation with R3D and I3D Backbones

In addition to the R(2+1)-18 backbone, we also show the performance of our proposed method with other commonly used video encoders *i.e.*, R3D-18 [24] and I3D [4]. For R3D-18, we use the same tubelet generation and transformation as that of R(2+1)D-18, as described in the main paper. For I3D, we change the input resolution to 224x224 and sample the patch size $H^{'} \times W^{'}$ uniformly from $[32{\times}32, 128{\times}128]$. For both, we follow the same pretraining protocol as described in the main paper.

We compare with prior works on the standard UCF101 [22] and HMDB51 [13] datasets. Table 2 shows the results with Kinetics-400 as the pretraining dataset. With the I3D backbone, our method outperforms prior works on both UCF101 and HMDB51. Similarly, with the R3D-18 backbone, we outperform prior works using the RGB modality on UCF101. We also achieve comparable performance to the best-performing method on HMDB51, improving over the next best method by 6.3%. On HMDB51 we also outperform prior works which pretrain on an additional optical flow modality and achieve competitive results with these methods on UCF101.

Method	**Modality**	UCF	HMDB
I3D			
SpeedNet [3]	RGB	66.7	43.7
DSM [26]	RGB	74.8	52.5
BE [27]	RGB	86.2	55.4
FAME [6]	RGB	88.6	61.1
This paper †	RGB	**89.5**	**64.0**
This paper	RGB	**89.7**	**63.9**
R3D-18			
VideoMoCo [17]	RGB	74.1	43.6
RSPNet [18]	RGB	74.3	41.6
LSFD [2]	RGB	77.2	53.7
MLFO [19]	RGB	79.1	47.6
ASCNet [10]	RGB	80.5	52.3
MCN [15]	RGB	85.4	54.8
TCLR [5]	RGB	85.4	55.4
CtP [25]	RGB	86.2	57.0
TE [11]	RGB	87.1	**63.6**
MSCL [16]	RGB+Flow	90.7	62.3
MaCLR [28]	RGB+Flow	91.3	62.1
This paper †	RGB	**88.8**	62.0
This paper	RGB	**90.1**	63.3

Table 2: **Evaluation with I3D and R3D backbones:** on standard UCF101 and HMDB51 benchmarks. Gray lines indicate the use of additional modalities during self-supervised pretraining. † indicates pretraining on Mini-Kinetics, otherwise, all models were pretrained on Kinetics-400.

C. Evaluation on Kinetics Dataset

To show whether our tubelet-contrastive pretraining can improve the performance of downstream tasks when plenty of labeled data is available for finetuning, we evaluate it on the Kinetics-400 [12] dataset for the task of action classification. The dataset contains about 220K labeled videos for training and 18K videos for validation. As evident from Table 4, such large-scale datasets can still benefit from our pretraining with a 3.4% improvement over training from scratch and 0.7% over the temporal contrast baseline.

D. Finetuning Details

During finetuning, we follow the setup from the SEVERE benchmark [23] which is detailed here for complete-

Evaluation Factor	Experiment	Dataset	Batch Size	Learning rate	Epochs	Steps
Standard	UCF101	UCF 101 [22]	32	0.0001	160	[60,100,140]
	HMDB51	HMDB 51 [13]	32	0.0001	160	[60,100,140]
Domain Shift	SS-v2	Something-Something [9]	32	0.0001	45	[25, 35, 40]
	Gym-99	FineGym [20]	32	0.0001	160	[60,100,140]
Sample Efficiency	UCF (10^3)	UCF 101 [22]	32	0.0001	160	[80,120,140]
	Gym (10^3)	FineGym [20]	32	0.0001	160	[80,120,140]
Action Granularity	FX-S1	FineGym [20]	32	0.0001	160	[70,120,140]
	UB-S1	FineGym [20]	32	0.0001	160	[70,120,140]
Task Shift	UCF-RC	UCFRep [29]	32	0.00005	100	-
	Charades	Charades [21]	16	0.0375	57	[41,49]

Table 3: **Training Details** of finetuning on various downstream datasets and tasks.

Pretraining	Top-1
None	61.4
Temporal Contrast Baseline	64.1
This paper	**64.8**

Table 4: **Kinetics-400 Evaluation.** Comparison with temporal contrastive pretraining for large-scale action recognition. All models use R(2+1)D-18 and pretraining was done on Kinetics-400 training set.

Transformation	UCF (10^3)	Gym (10^3)
None	63.0	45.6
Scale	65.1	46.5
Shear	65.2	47.5
Rotate	65.5	48.0
Scale + Shear	65.2	46.0
Rotate + Scale	65.4	46.9
Rotate + Shear	65.3	45.7
Rotate + Scale + Shear	65.6	46.0

Table 5: **Tubelet Transformation Combinations.** Combining transformations doesn't give a further increase in performance compared to using individual transformations.

ness. For all tasks, we replace the projection of the pre-trained model with a task-dependent head.

Action Recognition. Downstream settings which examine domain shift, sample efficiency, and action granularity all perform action recognition. We use a similar finetuning process for all experiments on these three factors. During the training process, a random clip of 32 frames is taken from each video and standard augmentations are applied: a multi-scale crop of 112x112 size, horizontal flipping, and color jittering. The Adam optimizer is used for training, with the learning rate, scheduling, and total number of epochs for each experiment shown in Table 3. During inference, 10 linearly spaced clips of 32 frames each are used, with a center crop of 112x112. To determine the action class prediction for a video, the predictions from each clip are averaged. For domain shift and sample efficiency, we report the top-1 accuracy. For action granularity experiments we report mean class accuracy, which we obtain by computing accuracy per action class and averaging over all action classes.

Repetition counting. The implementation follows the original repetition counting work proposed in UCFrep work [29]. From the annotated videos, 2M sequences of 32 frames with spatial size 112x112 are constructed. These are used as the input. The model is trained with a batch size of 32 for 100 epochs using the Adam optimizer with a learning rate of 0.00005. For testing, we report mean counting error following [29].

Multi-label classification on Charades. Following [8], a per-class sigmoid output is utilized for multi-class prediction. During the training process, 32 frames are sampled with a stride of 8. Frames are cropped to 112x112 and random short-side scaling, random spatial crop, and horizontal flip augmentations are applied. The model is trained for a total of 57 epochs with a batch size of 16 and a learning rate of 0.0375. A multi-step scheduler with $\gamma = 0.1$ is applied at epochs [41, 49]. During the testing phase, spatiotemporal max-pooling is performed over 10 clips for a single video. We report mean average precision (mAP) across all classes.

SSv2-Sub details. We use a subset of Something-Something v2 for ablations. In particular, we randomly sample 25% of the data from the whole train set and spanning all categories. This results in a subset consisting of 34409 training samples from 174 classes. We use the full validation set of Something-Something v2 for testing.

E. Tubelet Transformation Hyperparameters

Table 5 shows the results when applying multiple tubelet transformations in the tubelet generation. While applying individual transformations improves results, combing multiple transformations doesn't improve the performance further. This is likely because rotation motions are common in the downstream datasets while scaling and shearing are less common.

Min	Max	UCF (10^3)	Gym (10^3)
None			
-	-	63.0	45.6
Scale			
0.5	1.25	65.6	45.3
0.5	1.5	65.1	46.5
0.5	2.0	65.6	46.0
Shear			
-0.75	0.75	64.4	47.5
-1.0	1.0	65.2	48.0
-1.5	1.5	65.2	47.5
Rotation			
-45	45	65.2	49.3
-90	90	65.5	48.0
-180	180	65.6	49.6

Table 6: **Tubelet Transformation Hyperparameters.** We change Min and Max values for tubelet transformations. Our model is robust to changes in these parameters, with all choices tested giving an improvement over no tubelet transformation.

	UCF (10^3)	Gym (10^3)	SSv2-Sub	UB-S1
Randomly Scaled Crops	59.5	37.5	44.8	87.0
Tubelets	65.5	48.0	47.9	90.9

Table 7: **Tubelets vs Randomly Scaled Crops**. Our tubelets generate smooth motions to learn better video representations than strongly jittered crops.

Table 6 shows an ablation over Min and Max values for tubelet transformations. In the main paper, we use scale values between 0.5 and 1.5, shear values between -1.0 and 1.0, and rotation values between -90 and 90. Here, we experiment with values that result in more subtle and extreme variations of these transformations. We observe that all values for each of the transformations improve over no transformation. Our model is reasonably robust to these choices in hyperparameters, but subtle variations *e.g.*, scale change between 0.5 to 1.25 or shear from 0.75 to 0.75 tend to be slightly less effective.

F. Tubelets vs. Randomly Scaled Crops

To show that our proposed tubelets inject useful motions in the training pipeline, we compare them with randomly scaled crops. In particular, we randomly crop, scale, and jitter the patches pasted into the video clips when generating positive pairs and pretrain this and our model on Mini-Kinetics. Table 7 shows that our proposed motion tubelets outperform such randomly scaled crops in all downstream settings. This validates that the spatiotemporal continuity in motion tubelets is important to simulate smooth motions for learning better video representations.

G. Per-Class Results

Examining the improvement for individual classes gives us some insight into our model. Figure 1 shows the difference between our approach and the baseline for the 10 classes in UCF (10^3) with the highest increase and decrease in accuracy. Many of the actions that increase in accuracy are motion-focused, *e.g.*, pullups, lunges and jump rope. Other actions are confused by the baseline because of the similar background, *e.g.*, throw discus is confused with hammer throw and apply eye makeup is confused with haircut. The motion-focused features our model introduces help distinguish these classes. However, our model does lose some useful spatial features for distinguishing classes such as band marching and biking.

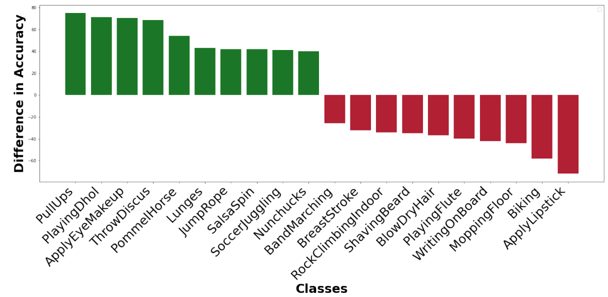

Figure 1: **Per-Class Accuracy Difference** on UCF (10^3) between our model and the temporal contrastive baseline. We show the 10 actions with the highest increase and decrease. Our model can better distinguish classes requiring motion but loses some ability to distinguish spatial classes.

H. Class Agnostic Activation Maps

Figure 2 show more examples of class agnostic activation maps [1] for video clips from various downstream datasets. Note that no finetuning is performed, we directly apply the representation from our tubelet contrastive learning pretrained on Kinetics-400. For examples from Fine-Gym, Something Something v2, and UCF101, we observe that our approach attends to regions with motion while the temporal contrastive baseline mostly attends to background.

I. Limitations and Future Work

There are several open avenues for future work based on the limitations of this work. First, while we compare to transformer-based approaches, we do not present the results of our tubelet-contrast with a transformer backbone. Our initial experiments with a transformer-based encoder [7] did not converge with off-the-shelf settings. We hope future work can address this problem for an encoder-independent solution. Additionally, we simulate tubelets with random image crops that can come from both background and foreground regions. Explicitly generating tubelets from foreground regions or pre-defined objects is a potential future direction worth investigating. Finally, we only simulate tubelets over short clips, it is also worth investigating whether long-range tubelets can be used for tasks that require long-range motion understanding.

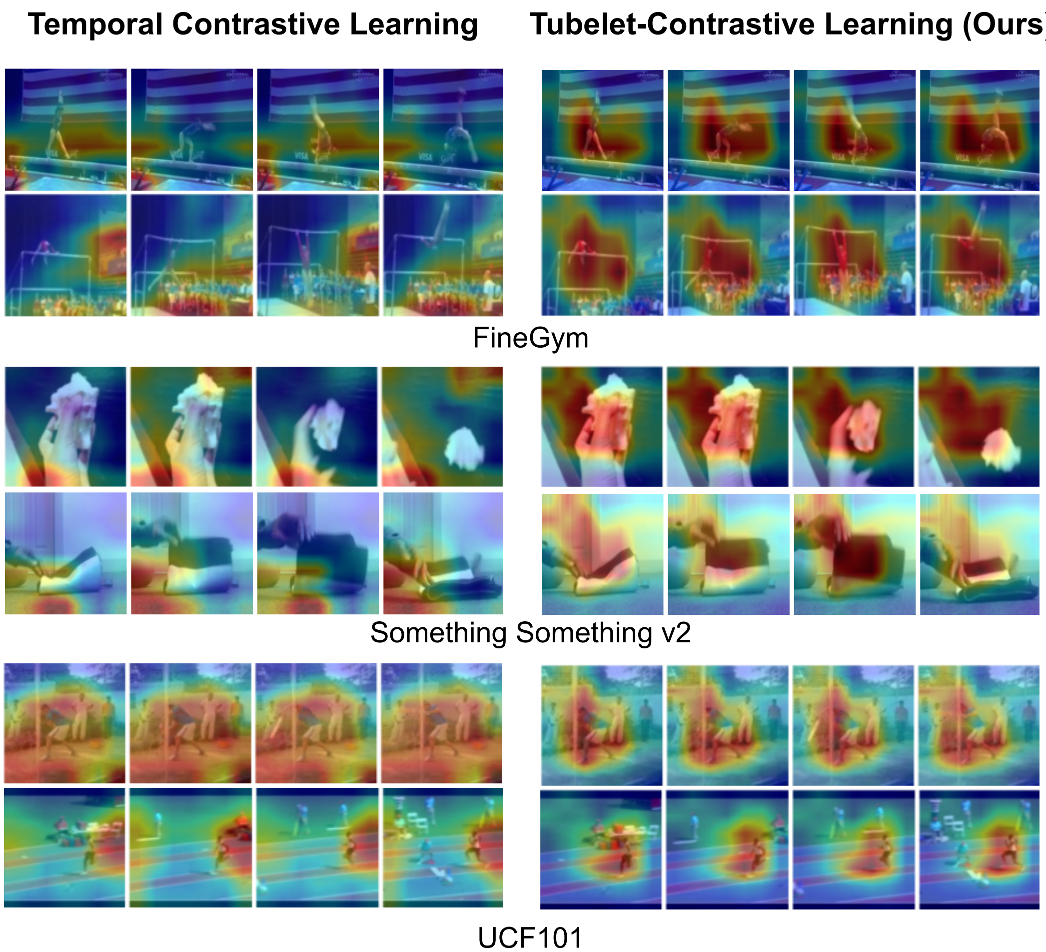

Figure 2: **Class-Agnostic Activation Maps Without Finetuning** for the temporal contrastive baseline and our tubelet contrast for different downstream datasets. Our model better attends to regions with motion irrespective of the domain.

References

[1] Kyungjune Baek, Minhyun Lee, and Hyunjung Shim. Psynet: Self-supervised approach to object localization using point symmetric transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 4

[2] Nadine Behrmann, Mohsen Fayyaz, Juergen Gall, and Mehdi Noroozi. Long short view feature decomposition via contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[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 (CVPR)*, 2020. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[5] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding (CVIU)*, 219:103406, 2022. 2

[6] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[8] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), 2021. 3

[9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[10] Deng Huang, Wenhao Wu, Weiwen Hu, Xu Liu, Dongliang He, Zhihua Wu, Xiangmiao Wu, Mingkui Tan, and Errui Ding. Ascnet: Self-supervised video representation learning with appearance-speed consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[11] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[12] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[13] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011. 2, 3

[14] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[15] Yuanze Lin, Xun Guo, and Yan Lu. Self-supervised video representation learning with meta-contrastive network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[16] Jingcheng Ni, Nan Zhou, Jie Qin, Qian Wu, Junqi Liu, Boxun Li, and Di Huang. Motion sensitive contrastive learning for self-supervised video representation. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[17] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[18] Chen Peihao, Huang Deng, He Dongliang, Long Xiang, Zeng Runhao, Wen Shilei, Tan Mingkui, and Gan Chuang. Rspnet: Relative speed perception for unsupervised video representation learning. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[19] Rui Qian, Yuxi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, and Weiyao Lin. Enhancing self-supervised video representation learning via multi-level feature optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[20] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[21] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 3

[23] Fida Mohammad Thoker, Hazel Doughty, Piyush Bagad, and Cees G M Snoek. How severe is benchmark-sensitivity in video self-supervised learning? In *European Conference on Computer Vision (ECCV)*, 2022. 2

[24] 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 (CVPR)*, 2018. 2

[25] Guangting Wang, Yizhou Zhou, Chong Luo, Wenxuan Xie, Wenjun Zeng, and Zhiwei Xiong. Unsupervised visual representation learning by tracking patches in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[26] Jinpeng Wang, Yuting Gao, Ke Li, Jianguo Hu, Xinyang Jiang, Xiaowei Guo, Rongrong Ji, and Xing Sun. Enhancing unsupervised video representation learning by decoupling the scene and the motion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[27] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[28] Fanyi Xiao, Joseph Tighe, and Davide Modolo. Maclr: Motion-aware contrastive learning of representations for videos. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[29] Huaidong Zhang, Xuemiao Xu, Guoqiang Han, and Shengfeng He. Context-aware and scale-insensitive temporal repetition counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3