Learning Video Representations without Natural Videos

Xueyang Yu Xinlei Chen Yossi Gandelsman

Abstract

We demonstrate that strong video representations can be learned without *natural videos. A simple progression of eight synthetic datasets—adding motion, acceleration, and shape/textural complexity—gradually boosts downstream accuracy: a VideoMAE model pre-trained on the final synthetic set closes* ***97.2 %*** *of the UCF-101 gap between scratch and natural-video SSL and* surpasses *that baseline on HMDB-51. Adding static ImageNet crops during pre-training fully matches UCF-101 performance and beats the UCF-101-pre-trained model on 11 / 14 UCF-101-P corruption variants. Our results show controllable, privacy-preserving synthetic data can rival large-scale natural videos for self-supervised learning*[1].

1. Introduction

Self-supervised pre-training on web-scale corpora has revolutionised NLP, yet comparable gains in video remain limited: even after ingesting millions of clips, current methods still trail supervised baselines on action recognition. We therefore ask a sharper question: *are real videos actually required to learn strong representations?*

To find out, we propose a progression of simple synthetic video generators that model a gradually growing set of video data properties – starting from static frames with solid-color circles and introducing additional shapes, dynamics, temporal shape changes, acceleration, and other textures. We show that adding each of the different properties improves the downstream video understanding performance; the final dataset—accelerating, transforming ImageNet crops—**matches or surpasses** UCF-101 pre-training and beats the UCF-101 baseline on 11/14 UCF-101-P corruptions.

By comparing the accuracy of models pre-trained on the generated data in the progression, we identify different data properties that correspond with improved downstream performance. We find that high velocities and accelerations of moving shapes in the video, as well as similarity in the color space to natural videos and high frame diversity, correlate to better action recognition accuracy. We believe that these observations can help to guide future practices for large-scale self-supervised video learning.

[1]Project page, code and data: https://unicorn53547.github.io/video_syn_rep/

2. Learning Video Representations without Natural Videos

To close the gap between training from scratch and natural video pre-training, we provide a progression of datasets. The datasets gradually introduce different aspects that appear in video data (e.g. transforming shapes, accelerating shapes). We provide the pre-training and downstream evaluation suit in appendix (Section 6.2).

2.1. Progression of video generation processes

We start by describing the progression of generative models $\{G_i\}$ we use to generate our training datasets. Each model uses a random number generator to sample latent parameters. The latent parameters are used for generating videos - sequences of T frames $f_t \in \mathbb{R}^{H\times W\times 3}, t \in \{1, ..., T\}$. Each consecutive model is built on top of the previous model, by modifying one aspect of it and adding additional calls to the random number generator. Examples of frames sampled from videos in the progression are shown in Figure 1. The models in the progression are described next (see Section 6.6 for additional hyper-parameters, and the supplementary material for videos).

Static circles. Our first video model is of static synthetic images of multiple circles that are copied T times (e.g. $f_t = f_{t+1}$). The color and location of the circles are sampled uniformly at random. Following the Dead Leaves [5], the radius is sampled from an exponential distribution.

Moving circles. Starting from randomly positioned circles in the first frame, each assigned a velocity to derive the next frames by modeling the dynamics. Each circle is assigned a random direction and a velocity magnitude that is sampled uniformly from a fixed range.

Moving shapes. We replace the circles sampled for the first frame with different shapes, including circles, quadrilaterals, and triangles. The shape types are sampled uniformly at random, and velocities are applied to them to simulate the next frames, similarly to the previous model.

Moving and transforming shapes. We introduce temporal transformations to the sampled shapes and apply them

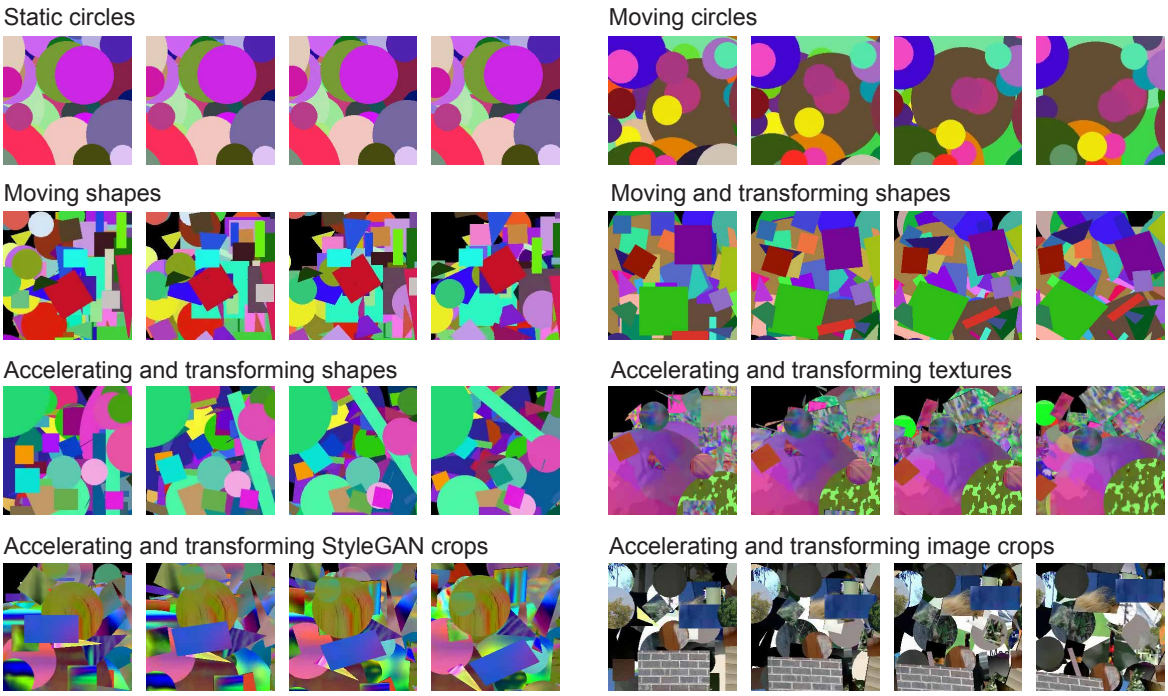

Figure 1. **Samples from our progression of video generation models and additionally included image datasets.** We present 4 frames from timestamps $t \in \{0, 10, 20, 30\}$ of a randomly sampled video from each of our generated datasets, and UCF101 (left to right).

together with the velocities to derive the next frames. Each shape is assigned two scaling factors (one for each spatial dimension), a rotation speed, and two sheer factors.

Accelerating transforming shapes. To introduce more complex dynamics, each temporally transforming shape is accelerated during the video by a random factor. The acceleration value is sampled uniformly from a fixed range that includes both positive and negative values.

Accelerating transforming textures. We replace the solid-colored shapes with textures from the statistical image dataset [2]. The dataset mimics color distribution, spectral components, and wavelet distribution of natural images and was shown to be useful for image pre-training.

Accelerating transforming StyleGAN crops. We replace the statistical textures with texture crops from the StyleGAN-Oriented dataset [2], which contains texture images that were sampled from an untrained StyleGAN [14] initialized to have the same wavelets for all output channels in the convolution layers.

Accelerating transforming image crops. We substitute the synthetic textures sampled for the previous Oriented-StyleGAN dataset with natural image crops, taken from ImageNet [7]. We do not parse or segment the images; instead, we sample random crops in the shapes mentioned above.

3. Experimental Results

We analyze how pre-training on data sampled from the generative models presented in Sec. 2 affects the downstream performance. We show results for fine-tuned models on in-distribution and out-of-distribution datasets (Sections 3.1 and 3.2) and for linear-probed models (Section 3.3).

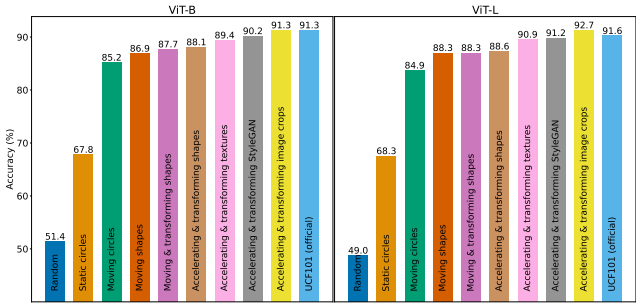

Figure 2. **Action recognition accuracy on UCF101.** We present the UCF101 classification accuracy of the progression of models $\{M_i\}$, after fine-tuning each of them on UCF101. The accuracy increases along the progression.

3.1. Fine-tuning

We fine-tune the pre-trained models for two different model scales, ViT-B and ViT-L, and evaluate the action recognition accuracy on UCF101, HMDB51, and Kinetics-400. We follow the protocol and hyper-parameters of [24] and tune only the learning rate and batch size.

UCF101 action classification. The results are presented in Figure 2. The final model in the progression, accelerating and transforming shapes with ImageNet crops, performs

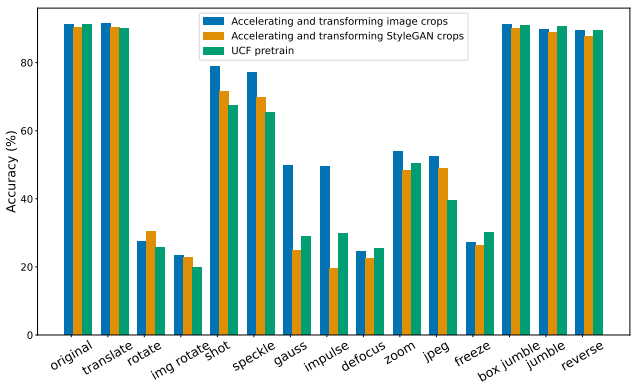

Figure 3. **Distribution Shift results on UCF101-P [20] (ViT-B)** The last model in our progression outperforms pre-training on natural videos for 11 out of 14 corruption datasets.

similarly to the model that was pre-trained on the UCF101 dataset (ViT-B), or outperforms it (ViT-L). Each fine-tuned model M_i in the progression improves over its predecessor for both model scales. A large increase in performance happens when dynamics are introduced to the generated data (e.g., from static circles to moving circles).

HMDB51 action classification. We evaluate the pre-trained models by fine-tuning them on the HMDB51 and present the results for ViT-B in Table 1. As shown, the order of the progression for the classification accuracy is similar. The two last models in our progression are more accurate than the model that was pre-trained on UCF101.

3.2. Distribution shift

We fine-tune the pre-trained models $\{M_i\}$ on UCF-101, and evaluate on corrupted datasets from UCF101-P [20]. The results for the last two models in the progression are presented in Figure 3. The last model in the progression outperforms the UCF101 pre-trained model on 11 out of 14 tasks and performs comparably on the rest. This suggests that the current pre-train recipe fails to generalize to out-of-distribution datasets. Additionally, the second to last model in our progression, which does not use real images, performs better only on 6 out of the 14 datasets. This suggests that differently from StyleGAN textures, the natural image crops unlock generalization capabilities to out-of-distribution *video* corruptions.

3.3. Linear-probing

We probe the progression of pre-trained models on UCF101. The results are presented in Table 1. The difference in performance between the last model in the progression and the model trained on UCF101 is more significant (a gap of 23.2%). Compared to the best model of [2] that was trained on synthetic image data, which closes 56.5% of the gap between linear probing on randomly initialized weights and linear probing on a pre-trained model, the last model in our progression closes only 40.6% of the gap. We suspect

	HMDB51 fine-tune	UCF101 lin. prob	UCF101 fine-tune
Random initialization	18.2	8.9	51.4
Static circles	29.2	13.2	67.8
Moving circles	52.0	15.5	85.2
Moving shapes	56.1	20.4	86.9
Moving and transforming shapes	57.6	18.8	87.7
Acc. and transforming shapes	58.9	18.9	88.1
Acc. and transforming textures	62.4	20.9	89.4
Acc. and transforming StyleGAN crops	**64.1**	25.2	90.2
Acc. and transforming image crops	**64.1**	24.8	**91.3**
UCF101	63.0	**48.0**	**91.3**

Table 1. **Additional action recognition results (ViT-B).** We present the classification accuracy on HMDB51 after fine-tuning and on UCF101 after linear probing/fine-tuning for all the pre-training datasets in our progression and the two baselines.

that the difference in the gap between fine-tuning and linear probing is due to large differences between low-level characteristics of natural images and our datasets, which can be mitigated by fine-tuning the full model. We analyze these low-level properties in Section 4.3.

4. Datasets Analysis

We analyze in depth different characteristics of the synthetic datasets that were shown to be useful for video pre-training. We start by evaluating the effect of incorporating natural images during training. Then, we analyze the effects of different types of synthetic textures. Finally, we compare the videos statistical properties to downstream performance.

4.1. Incorporating static images

Following the improvement when natural image crops are used, we ask: 1) how does the size of the static image affect the downstream performance, 2) can the pre-training benefit from both synthetic and natural crops.

Image dataset size. We evaluate the effect of the image data size on the downstream task. Our initial pool of images includes all the images from ImageNet (1.3M). We provide additional results with 300k images *while keeping the size of the pre-training video dataset fixed*. The results are presented in Table 2. An increase in the static images dataset results in a better performance on the downstream task.

Combining natural images and synthetic textures. To evaluate if useful pre-training can be achieved by combining natural images and synthetic textures, we create a dataset that incorporates crops from half of the images and crops from half of the synthetic textures from the StyleGAN textures [2]. As shown in Table 2, the performance of the new dataset ("150k images & 150k StyleGAN") is slightly higher than the performance of the two datasets that use solely one type of data. This suggests that mixing datasets can lead to improved performance in other cases as well. We leave this approach to future work.

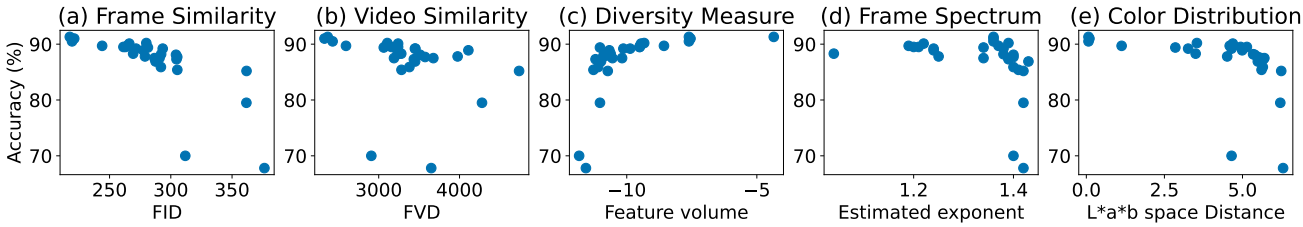

Figure 4. **Dataset properties compared to downstream performance.** We compare the downstream classification accuracy on UCF101 after fine-tuning to frame and video properties of all the dataset variants we used in our analysis (see datasets list in Section 6.6).

Configuration	Accuracy (%)
300k images	90.5
150k images & 150k StyleGAN	90.6
300k StyleGAN	90.2
300k statistical textures	89.4
1.3M images	91.3

Table 2. **Incorporating natural images into training (ViT-B).** We ablate different approaches for incorporating natural images during training, and evaluate them on UCF101.

Configuration (ViT-B)	Accuracy (%)
Static StyleGAN crops	90.2
Dynamic StyleGAN crops	89.2
Dynamic StyleGAN videos	68.7

Table 3. **Using synthetic textures during training.** Introducing dynamic StyleGAN textures does not improve performance.

4.2. Incorporating dynamic textures

We evaluate additional synthetic dynamics that can be incorporated in our progression. Specifically, we replace the static StyleGAN textures with a dynamic version.

Dynamic StyleGAN textures. Starting from a latent z_0, each frame is rendered as $G'(z_i)$ where $z_i = z_{i-1} + \delta z_i$ and $\delta z_i \sim \mathcal{N}(0, \sigma^2)$. We evaluate two uses: (i) standalone clips (*Dynamic StyleGAN videos*) and (ii) replacing the solid-colour shapes in the "accelerating + transforming shapes" set with frame-wise updated texture crops (*Dynamic StyleGAN crops*).

UCF101 fine-tune. Table 3 presents the action classification accuracy after incorporating the dynamic textures in the pre-training stage and fine-tuning on UCF101. Both results suggest that the simple hand-crafted dynamics of randomly moving Dead-Leaves models are sufficient for pre-training, without additional dynamics modeling.

4.3. Similarity and static property analysis

During our experiments, we generated 28 datasets and trained ViT-B VideoMAE on each (see Section 6.8). We plot the models' UCF101 fine-tuning accuracies as a function of their similarity to UCF101. Following [2], we also analyze low-level statistics of generated data.

FID. As shown in Figure 4.a There is a strong negative correlation between the frame similarity to the accuracy ($r = -0.72$). This suggests that improving frame similarity can lead to better performance.

FVD. Differently from the frame similarity, there is less significant negative correlation between the FVD metric and the performance ($r = -0.27$). This suggests that this metric is less indicative of downstream performance.

Diversity. We utilize inception features [23] and plot the determinant of their covariance matrix. As shown in Figure 4.c, the datasets that include synthetic textures and image crops are more diverse than the other datasets. This suggests that investing in more diverse datasets can improve performance even further.

Image spectrum. Following [25], that showed that the spectrum of natural images resembles the function $A/|f|^{\alpha}$, with an exponent α ranging in $[0.5, 2.0]$, we estimate the exponent for frames in our datasets. The datasets that result in the best downstream performance have an α that lies close to the middle of the range.

Color statistics. We model the color distributions as three-dimensional Gaussian that correspond to the three color channels in L*a*b space and compute the symmetric KL divergence between the color distributions of each dataset.There is a relatively weak negative correlation of $r = -0.42$ between the color distance to UCF101 and the accuracy.

5. Discussion

Learning from *synthetic* data offers a key advantage: full control over content, eliminating the risk of hidden malicious, private, or biased samples that plague web-scale video corpora. In this paper, we show that simple, well-understood generators already rival natural-video pre-training, suggesting a principled path toward ever better synthetic datasets. We believe that our synthetic data study can be utilized to create better datasets for learning video representations without natural videos. Guided by the analysis, we plan to explore other well-understood data sources and generation processes to continue improving video representation learning, in large-scale training regimes.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. On the effectiveness of vit features as local semantic descriptors. In *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, page 39–55, Berlin, Heidelberg, 2023. Springer-Verlag. 2

[2] Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. In *Advances in Neural Information Processing Systems*, 2021. 2, 3, 4, 1, 5

[3] Manel Baradad, Chun-Fu Chen, Jonas Wulff, Tongzhou Wang, Rogerio Feris, Antonio Torralba, and Phillip Isola. Procedural image programs for representation learning. In *Advances in Neural Information Processing Systems*, 2022. 1

[4] 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. 1

[5] Charles Bordenave, Yann Gousseau, and François Roueff. The dead leaves model: A general tessellation modeling occlusion. *Advances in Applied Probability*, 38(1):31–46, 2006. 1

[6] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[8] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[9] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[10] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 1

[11] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[12] Xi Guo, Wei Wu, Dongliang Wang, Jing Su, Haisheng Su, Weihao Gan, Jian Huang, and Qin Yang. Learning video representations of human motion from synthetic data. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20165–20175, 2022. 1

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings - 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 15979–15988. IEEE Computer Society, 2022. Publisher Copyright: © 2022 IEEE.; 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022 ; Conference date: 19-06-2022 Through 24-06-2022. 1

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 2

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[16] YoWhan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris. How transferable are video representations based on synthetic data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 2

[18] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *CoRR*, abs/1511.05440, 2015. 1

[19] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 527–544. Springer, 2016. 1

[20] Madeline C Schiappa, Naman Biyani, Prudvi Kamtam, Shruti Vyas, Hamid Palangi, Vibhav Vineet, and Yogesh Rawat. Large-scale robustness analysis of video action recognition models. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3, 2

[21] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, page 568–576, Cambridge, MA, USA, 2014. MIT Press. 1

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

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. 4

[24] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 2, 1

[25] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3): 391–412, 2003. PMID: 12938764. 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] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023. 1

[28] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 1

[29] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 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 *Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[31] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023. 1