Non-Adversarial Video Synthesis with Learned Priors
(Supplementary Material)

Abhishek Aich[†,*], Akash Gupta[†,*], Rameswar Panda[‡], Rakib Hyder[†],
M. Salman Asif[†], Amit K. Roy-Chowdhury[†]
University of California, Riverside[†], IBM Research AI, Cambridge[‡]
{aaich001@, agupt013@, rpand002, rhyde001@, sasif@ece, amitrc@ece}.ucr.edu

Table 1: Supplementary Material Overview.

[*]Joint first authors

A. Dataset Descriptions

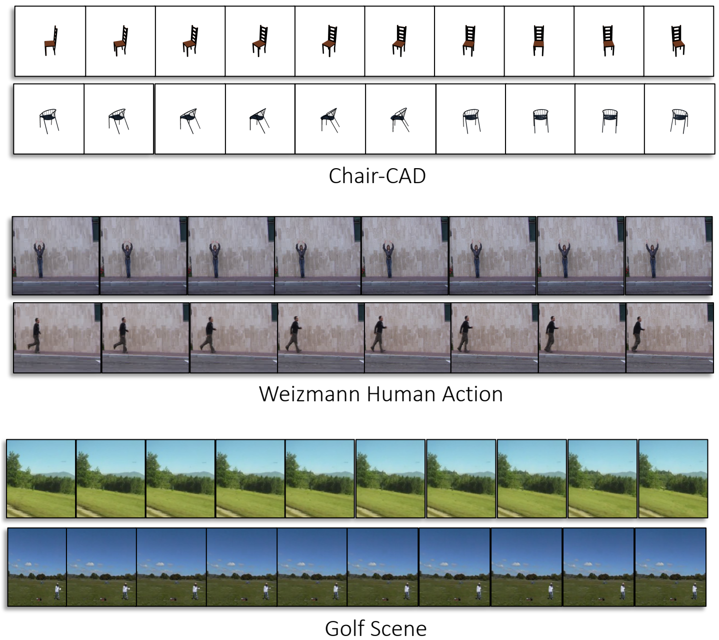


Figure 1: **Sample videos from datasets used in the paper.** Two unprocessed video examples from Chair-CAD [1], Weizmann Human Action [4], and Golf Scene [11] datasets have been presented here. As seen from the examples, datasets are diverse in nature, different in categories and present unique challenges in learning the transient and static portions of the videos. Best viewed in color.

Chair-CAD [1]**.** This dataset provides 1393 chair-CAD models. Each model frame sequence is produced using two elevation angles in addition to thirty one azimuth angles. All the chair models have been designed to be at a fixed distance with respect to the camera. The authors provide four video sequences per CAD model. We choose the first 16 frames of each video for our paper, and consider the complete dataset as one class.

Weizmann Human Action [2]**.** This dataset is a collection of 90 video sequences showing nine different identities performing 10 different actions, namely, `run`, `walk`, `skip`, `jumping-jack` (or 'jack'), `jump-forward-on-two-legs` (or 'jump'), `jump-in-place-on-two-legs` (or 'pjump'), `gallopsideways` (or 'side'), `wave-two-hands` (or 'wave2'), `waveone-hand` (or 'wave1'), and `bend`. We randomly choose 16 consecutive frames for every video in each iteration during training.

Golf Scene [11]. [11] released a dataset containing 35 million clips (32 frames each) stabilized by SIFT+RANSAC. It contains several categories filtered by a pre-trained Place-CNN model, one of them being the Golf scenes. The Golf scene dataset contains 20,268 golf videos. Due to many non-golf videos being part of the golf category (due to inaccurate labels), this dataset presents a particularly challenging data distribution for our proposed method. Note that for a fair comparison, we further selected our training set videos from this provided dataset pertaining to golf action as close as possible. We then trained the VGAN [11] model on this selected videos for a fair comparison.

B. Implementation Details

We used Pytorch [5] for our implementation. The Adam optimizer [3], with $\epsilon = 10^{-8}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, was used to update the model weights and SGD optimizer [6], with momentum $= 0.9$, was used to update the latent spaces. The corresponding learning rate for the generator $\tau_{\mathcal{G}}$, the RNN $\tau_{\mathcal{R}}$, and the latent spaces $\tau_{\mathcal{Z}_\mathrm{v}}$ were set as values indicated in Tab. 2.

Hyper-parameters. [7, 10, 11] that generate videos from latent priors have no dataset split as the task is to synthesize high quality videos from the data distribution, and then evaluate the model performance. All hyperparameters (except D_s, D_t) are set as described in [8, 9, 10, 12] (e.g. α from [9]). For D_s and D_t, we follow the strategy used in Sec. 4.3 of [10] and observe that our model generates videos with good visual quality (FCS) and plausible motion (MCS) for Chair-CAD when (D_s, D_t) = (206, 50). Same strategy is used for all datasets. The hyper-parameters employed with respect to each dataset used in this paper is given in Tab. 2. $\mathcal{G}$ and $\mathcal{R}$ refer to the generator, with weights γ, and RNN, with weight θ, respectively. $\tau_{(\cdot)}$ represents the learning rate. $\mu_{(\cdot)}$ represents the number of epochs. D_{s} and D_{t} refer to the static and transient latent dimensions, respectively. λ_{s} and λ_{t} refer to the static loss, and triplet loss regularization constants, respectively. α is the margin for triplet loss. l refers to the level of the Laplacian pyramid representation used in ℓ_{rec} and ℓ_{static}.

Datasets	Hyper-parameters										
	D_{s}	D_{t}	λ_{s}	λ_{t}	α	$\tau_{\mathcal{G}}$	$\tau_{\mathcal{R}}$	$\tau_{\mathcal{Z}_{\mathsf{V}}}$	μ_{γ}	$\mu_{\mathcal{Z}_{\mathsf{V}},\theta}$	l
Chair-CAD [1]	206	50	0.01	0.01	2	6.25×10^{-5}	6.25×10^{-5}	12.5	5	300	4
Weizmann Human Action [2]	56	200	0.01	0.1	2	6.25×10^{-5}	6.25×10^{-3}	12.5	5	700	3
Golf Scene [11]	56	200	0.01	0.01	2	0.1	0.1	12.5	10	1000	4

Table 2: Hyper-parameters used in all experiments for all datasets.

Other details. We performed all our experiments on a system with 48 core Intel(R) Xeon(R) Gold 6126 processor with 256GB RAM. We used NVIDIA GeForce RTX 2080 Ti for all GPU computations during training. Further, NVIDIA Tesla K40 GPUs were used for computation of all evaluation metrics in our experiments. All our implementations are based on non-optimized PyTorch based codes. Our runtime analysis revealed that it took on average one to two days to train the model and obtain learned latent vectors.

C. More Qualitative Examples

In this section, we provide more qualitative results of generated videos synthesized using our proposed approach on each dataset (Fig. 2 for Chair-CAD [1] dataset, Fig. 3 for Weizmann Human Action [2] dataset, and Fig. 4 for Golf scene [11] dataset). We also provide more examples interpolation experiment in Fig. 5.

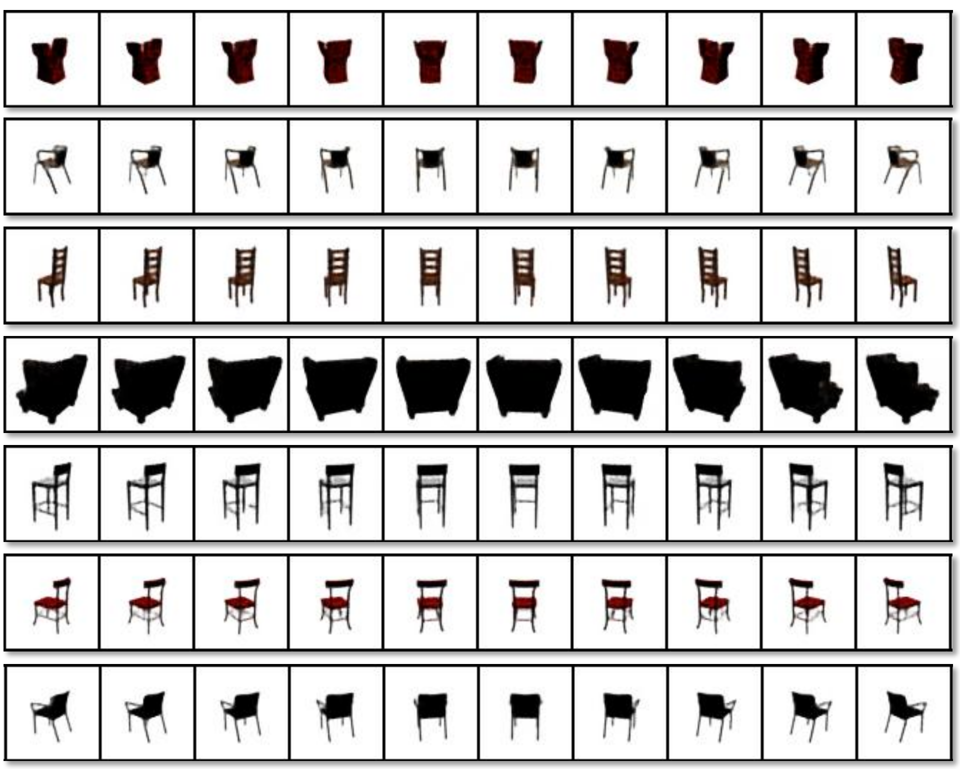
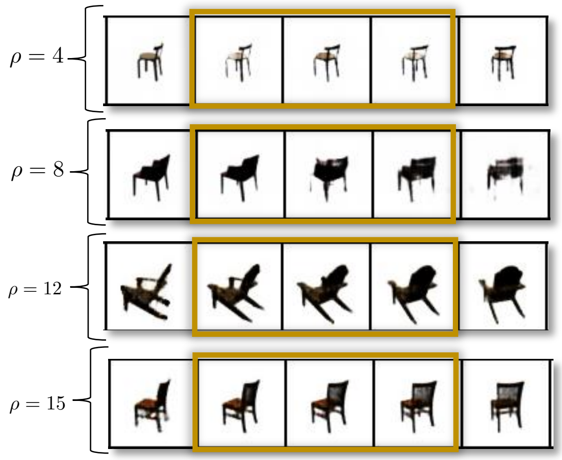


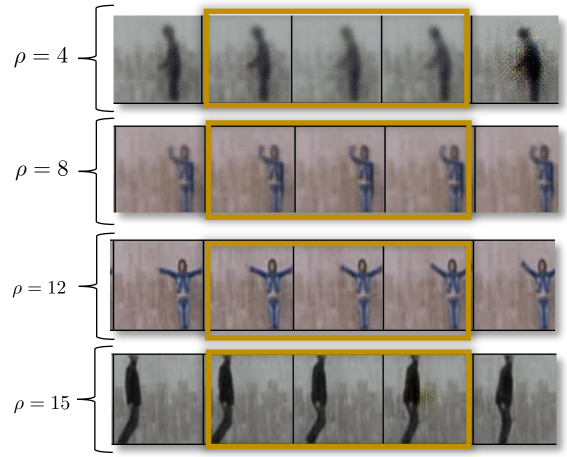
Figure 2: **Qualitative results on Chair-CAD** [1]. On this large scale dataset, our model is able to capture the intrinsic rotation and color of videos unique to each chair model. This shows the efficacy of our approach, compared to adversarial approaches such as MoCoGAN [10] which produce the same chair for all videos, with blurry frames (See Fig. 1 of main manuscript).

Figure 3: **Qualitative results on Weizmann Human Action** [2]. The videos show that our model produces sharp visual results with the combination of trained generator, RNN along with 9 identities, and 10 different action latent vectors.

Figure 4: **Qualitative results on Golf Scene** [11]. Our proposed approach produces visually good results on this particularly challenging dataset. Due to incorrect labels on the videos, this dataset has many non-golf videos. Our model is still able to capture the static and transient portion of the videos, although better filtering can still improve our results.



(a) Chair-CAD [1]



(b) Weizmann Human Action [2]

Figure 5: **More interpolation results**. In this figure, ρ represents the rank of transient latent vectors $\mathbf{z}_t$. We present the interpolation results on (a) Chair-CAD dataset, and (b) Weizmann Human Action, for different values of ρ. It can be observed that as ρ increases, the interpolation becomes clearer.

References

[1] Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell, and Josef Sivic. Seeing 3D Chairs: Exemplar part-based 2D-3D Alignment using a Large Dataset of CAD Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3769, 2014. 1, 2, 3, 5

[2] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as Space-Time Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007. 1, 2, 3, 4, 5

[3] Diederik P Kingma and Jimmy Ba. ADAM: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[4] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Networks. *arXiv preprint arXiv:1411.1784*, 2014. 2

[5] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic Differentiation in PyTorch. In *NIPS AutoDiff Workshop*, 2017. 2

[6] Sebastian Ruder. An Overview of Gradient Descent Optimization Algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 2

[7] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal Generative Adversarial Nets with Singular Value Clipping. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2830–2839, 2017. 2

[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FACENET: A Unified Embedding for Face Recognition and Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 2

[9] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-Contrastive Networks: Self-Supervised Learning from Video. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1134–1141, 2018. 2

[10] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MOCOGAN: Decomposing Motion and Content for Video Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2018. 2, 3

[11] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating Videos with Scene Dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016. 1, 2, 3, 4

[12] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems*, pages 1144–1156, 2018. 2