Non-Adversarial Video Synthesis with Learned Priors

Abhishek Aich[†,*], **Akash Gupta**[†,*], **Rameswar Panda**[‡], **Rakib Hyder**[†], **M. Salman Asif**[†], **Amit K. Roy-Chowdhury**[†]

University of California, Riverside[†], IBM Research AI, Cambridge[‡]

UC RIVERSIDE UNIVERSITY OF CALIFORNIA IBM **Research** AI CVPR SEATTLE WASHINGTON JUNE 16-18 2020

[*] Joint first authors

Outline

Problem Overview

Brief Statement:
How can we synthesize videos from random noise vectors without visual cues?

Why is the problem important?

- Understanding how videos are generated can help us in their decomposition of spatial and temporal behavior.
- Can be used as the building block for choosing and designing priors for different problems.
- Will help us in difficult problems like future frame prediction, feature learning for videos, etc.

Related Works

Prior works in this problem

- VGAN [1] demonstrates that a video can be divided into foreground and background using deep neural networks.
- TGAN [2] proposes to use a generator to capture temporal dynamics by generating correlated latent codes for each video frame and then using an image generator to map each of this latent code to a single frame for the whole video.
- MoCoGAN [3] presents a simple approach to separate content and motion latent codes of a video using adversarial learning.
- Video-VAE [4] extends the idea of image generation to video generation using Variational Auto-Encoder (VAE) by proposing a structured latent space in conjunction with the VAE architecture for videos synthesis using a given/generated input frame.

Related Works

Methods	Settings		
	Adversarial learning?	Input frame?	Input latent vectors?
VGAN [1]	✓	✗	✓(random)
TGAN [2]	✓	✗	✓(random)
MoCoGAN [3]	✓	✗	✓(random)
Video-VAE [4]	✗	✓	✓(random)
Ours	✗	✗	✓(learned)

Table 1: **Categorization of prior works in video synthesis.** Different from existing methods, our model doesn't require a discriminator, or any reference input frame. However, since we have learned latent vectors, we have control of the kind of videos the model should generate.

Proposed Approach: Motivation

Adversarial approaches [1–3] involve training a generative network using a discriminator. These are great for learning complex data-distribution, but come with following drawbacks:

- They are difficult to train given the saddle-point based learning surface.
- They suffer from mode collapse problem.
- They suffer from vanishing gradient problem due to the adversarial process.

Also, inspired from ···

Non-adversarial learning of generative networks for image generation [5, 6] have shown that properties of GANs (convolutional networks) can be mimicked using simple reconstruction losses while discarding the discriminator.

Proposed Approach: Notation

Define a video clip V represented by L frames as $\mathrm{V} = \left[v_1, v_2, \cdots, v_L\right]$. Corresponding to each frame, let there be a point in latent space $\mathcal{Z}_\mathrm{V} \in \mathbb{R}^{D \times L}$ such that

$$\mathcal{Z}_\mathrm{V} = \left[z_1, z_2, \cdots, z_L\right] \tag{1}$$

- We propose to disentangle a video into two parts: a static constituent, which captures the constant portion of the video common for all frames, and a transient constituent which represents the temporal dynamics between all the frames in the video.
- Assuming that the video is of short length, we can fix $z_i^{(\mathrm{s})} = z^{(\mathrm{s})}$ for all frames after sampling only once. Therefore, (1) can be expressed as

$$\mathcal{Z}_\mathrm{V} = \left[\begin{bmatrix} z^{(\mathrm{s})} \\ z_1^{(\mathrm{t})} \end{bmatrix}, \begin{bmatrix} z^{(\mathrm{s})} \\ z_2^{(\mathrm{t})} \end{bmatrix}, \cdots, \begin{bmatrix} z^{(\mathrm{s})} \\ z_L^{(\mathrm{t})} \end{bmatrix}\right] \tag{2}$$

Proposed Architecture

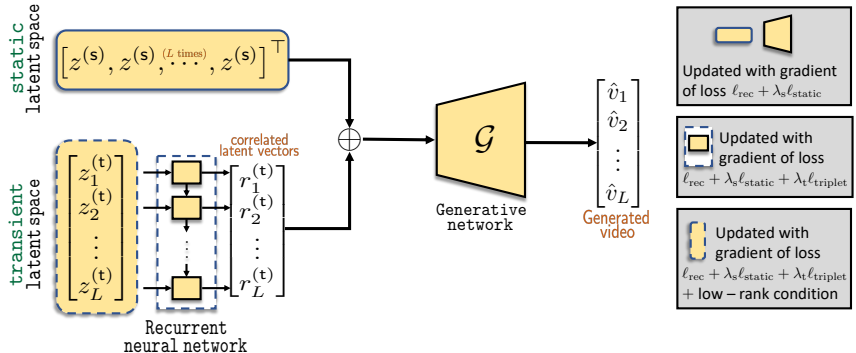


Figure 1: **Overview of Proposed Architecture.** We map a video (with L frame sequence) into two learnable latent spaces. We jointly learn the static latent space and the transient latent space along with the network weights.

Learning Loss I

Learning the weights of the network:

Let weights of $\mathcal{G}$, and the RNN be γ and θ, respectively. We jointly optimize for θ, γ, and $\{\mathcal{Z}_{\mathsf{V}_j}\}_{j=1}^{N}$ (sampled once in the beginning of training) for every epoch in two stages:

$$\text{Stage 1}: \quad \min_{\gamma} \ \ell\big(\mathsf{V}_j, \mathcal{G}(\mathcal{Z}_{\mathsf{V}_j})|\big(\mathcal{Z}_{\mathsf{V}_j}, \theta\big)\big) \tag{3.1}$$

$$\text{Stage 2}: \quad \min_{\mathcal{Z}_{\mathsf{V}}, \theta} \ \ell\big(\mathsf{V}_j, \mathcal{G}(\mathcal{Z}_{\mathsf{V}_j})|\gamma\big) \tag{3.2}$$

The index j represents a random video out of N videos chosen from the dataset. $\ell(\cdot)$ can be chosen to be any distance based loss. We will refer to both (3.1) and (3.2) together as $\min_{\mathcal{Z}_{\mathsf{V}}, \theta, \gamma} \ell_{\mathsf{rec}}$.

Learning Loss II

To equally capture the static portion of the video, we randomly choose a frame from the video and ask the generator to compare its corresponding generated frame during training. For this, we update the above loss as follows.

$$\min_{\mathcal{Z}_V,\theta,\gamma} \left(\ell_{\text{rec}} + \lambda_s \ell_{\text{static}}\right) \tag{4}$$

where $\ell_{\text{static}} = \ell(\hat{v}_k, v_k)$ with $k \in [1, 2, \cdots, L]$ is a randomly chosen index, v_k is the ground truth frame, $\hat{v}_k = \mathcal{G}(\mathbf{z}_k)$, and λ_s is the regularization constant.

Triplet Condition

Learning the latent space:

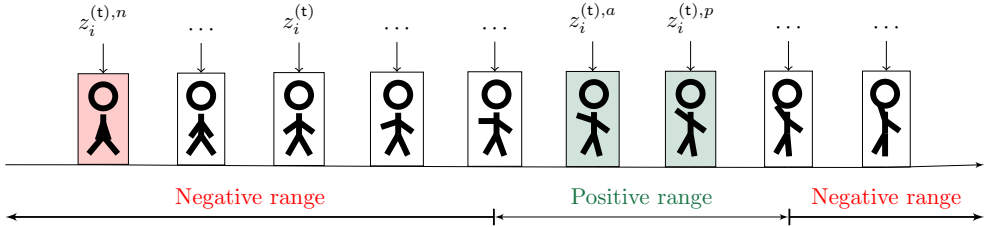


Figure 2: **Triplet Condition in the transient latent space.** Latent code representation of short video clips may lie very near to each other in the transient subspace. Using the proposed triplet condition, our model learns to explain the dynamics of similar looking frames and map them to distinct latent vectors.

Triplet Condition

Positive frames are randomly sampled within a margin range α of the anchor and negatives are chosen outside of this margin range. Defining a triplet set with transient latent code vectors $\{z_i^{(t),a}, z_i^{(t),p}, z_i^{(t),n}\}$, we aim to learn the transient embedding space $\mathbf{z}^{(t)}$ such that

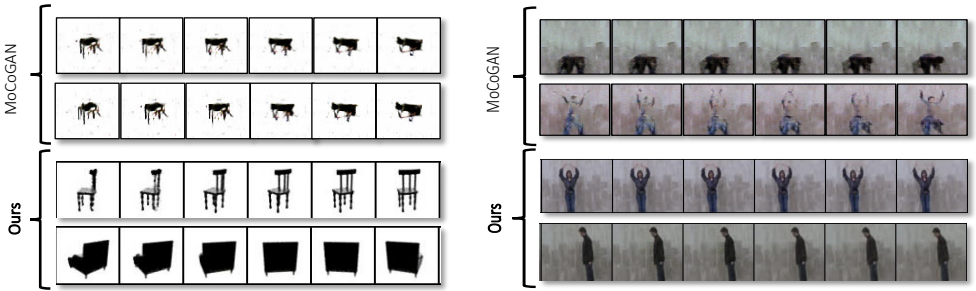
$$\|z_i^{(t),a} - z_i^{(t),p}\|_2^2 + \alpha < \|z_i^{(t),a} - z_i^{(t),n}\|_2^2$$

$\forall \{z_i^{(t),a}, z_i^{(t),p}, z_i^{(t),n}\} \in \Gamma$, where Γ is the set of all possible triplets in $\mathbf{z}^{(t)}$. With the above regularization, the loss in (4) can be written as

$$\min_{\mathcal{Z}_v, \theta, \gamma} \; (\ell_{\text{rec}} + \lambda_s \ell_{\text{static}})$$
$$\text{s.t.} \quad \|z_i^{(t),a} - z_i^{(t),p}\|_2^2 + \alpha < \|z_i^{(t),a} - z_i^{(t),n}\|_2^2 \tag{5}$$

where α is a hyperparameter that controls the margin while selecting positives and negative.

Qualitative Results



(a) Chair-CAD [7] (b) Weizmann Human Action [8]

Figure 3: **Qualitative results.** The proposed method produces visually sharper, and consistently better using the non-adversarial training protocol.

Quantitative Results

	MCS ↓	FCS ↑
Bound	0.0	0.91
MoCoGAN [3]	4.11	0.85
Ours ($-\ell_{\text{triplet}}-\ell_{\text{static}}$)	3.83	0.77
Ours ($+\ell_{\text{triplet}}+\ell_{\text{static}}$)	**3.32**	**0.89**

(a) Chair-CAD [7]

	MCS ↓	FCS ↑
Bound	0.0	0.95
MoCoGAN [3]	3.41	0.85
Ours ($-\ell_{\text{triplet}}-\ell_{\text{static}}$)	3.87	0.79
Ours ($+\ell_{\text{triplet}}+\ell_{\text{static}}$)	**2.63**	**0.90**

(b) Weizmann Human Action [8]

Table 2: **Quantitative results.** We obtained better scores on the proposed method on both Chair-CAD [7], and Weizmann Human Action [8] datasets, compared to the adversarial approaches (MoCoGAN, and VGAN[1]). Best scores have been highlighted in bold.

[1] not shown here

Application: Action exchange

Actions	Identities								
	P1	P2	P3	P4	P5	P6	P7	P8	P9
run	●	●	●		●				
walk	●	●			●		●		
jump		●		●	●	●	●	●	●
skip	●	●	●			●	●	●	●

Table 3: **Generating videos by exchanging unseen actions by identities.** Each cell in this table indicates a video in the dataset. Only cells containing the symbol ● indicate that the video was part of the training set. We randomly generated videos corresponding to rest of the cells indicated by symbols ●, ●, ●, and ●, visualized in Fig. 4.

Application: Action exchange

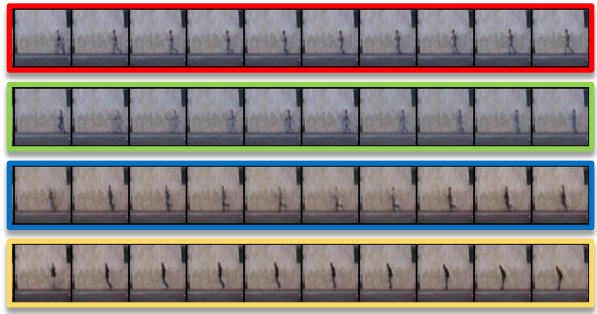


Figure 4: **Examples of action exchange to generate unseen videos.** This figure demonstrates the effectiveness of our method in disentangling static and transient portion of a video, as well as the efficacy in generating videos unseen during training. The colored bounding boxes indicate the unseen video generated referred in Tab. 3.

Conclusions

- We presented a non-adversarial approach for synthesizing videos by jointly optimizing both network weights and input latent space.
- Our approach allows us to generate videos from any mode of data distribution, perform frame interpolation[2], and generate videos unseen during training.
- Experiments on three[3] standard datasets show the efficacy of our proposed approach over state-of-the-methods.

Acknowledgment

This work was partially supported by NSF grant 1664172 & ONR grant N00014-19-1-2264.

[2] not shown in slides
[3] only two shown in slides

References I

[1] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics". In: *Advances In Neural Information Processing Systems*. 2016, pp. 613–621.

[2] 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*. 2017, pp. 2830–2839.

[3] Sergey Tulyakov et al. "Mocogan: Decomposing motion and content for video generation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1526–1535.

[4] Jiawei He et al. "Probabilistic video generation using holistic attribute control". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 452–467.

References II

[5] Piotr Bojanowski et al. "Optimizing the Latent Space of Generative Networks". In: *International Conference on Machine Learning*. 2018, pp. 599–608.

[6] Yedid Hoshen, Ke Li, and Jitendra Malik. "Non-Adversarial Image Synthesis with Generative Latent Nearest Neighbors". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5811–5819.

[7] Mathieu Aubry et al. "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*. 2014, pp. 3762–3769.

[8] Lena Gorelick et al. "Actions as space-time shapes". In: *IEEE transactions on pattern analysis and machine intelligence* 29.12 (2007), pp. 2247–2253.

Thank you!