UC RIVERSIDE UNIVERSITY OF CALIFORNIA IBM Research AI

# 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[‡]

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

## Brief Overview

Despite their success, existing works in video synthesis [1–3] often require input reference frame or fail to generate diverse videos from the given data distribution, with little to no uniformity in the quality of generated videos. Different from such methods and inspired from [4, 5], we focus on the problem of generating videos from latent noise vectors, without any reference input frames. To this end, we develop a novel approach that jointly optimizes the input latent space, the weights of a recurrent neural network and a generator without a discriminator, through non-adversarial learning.

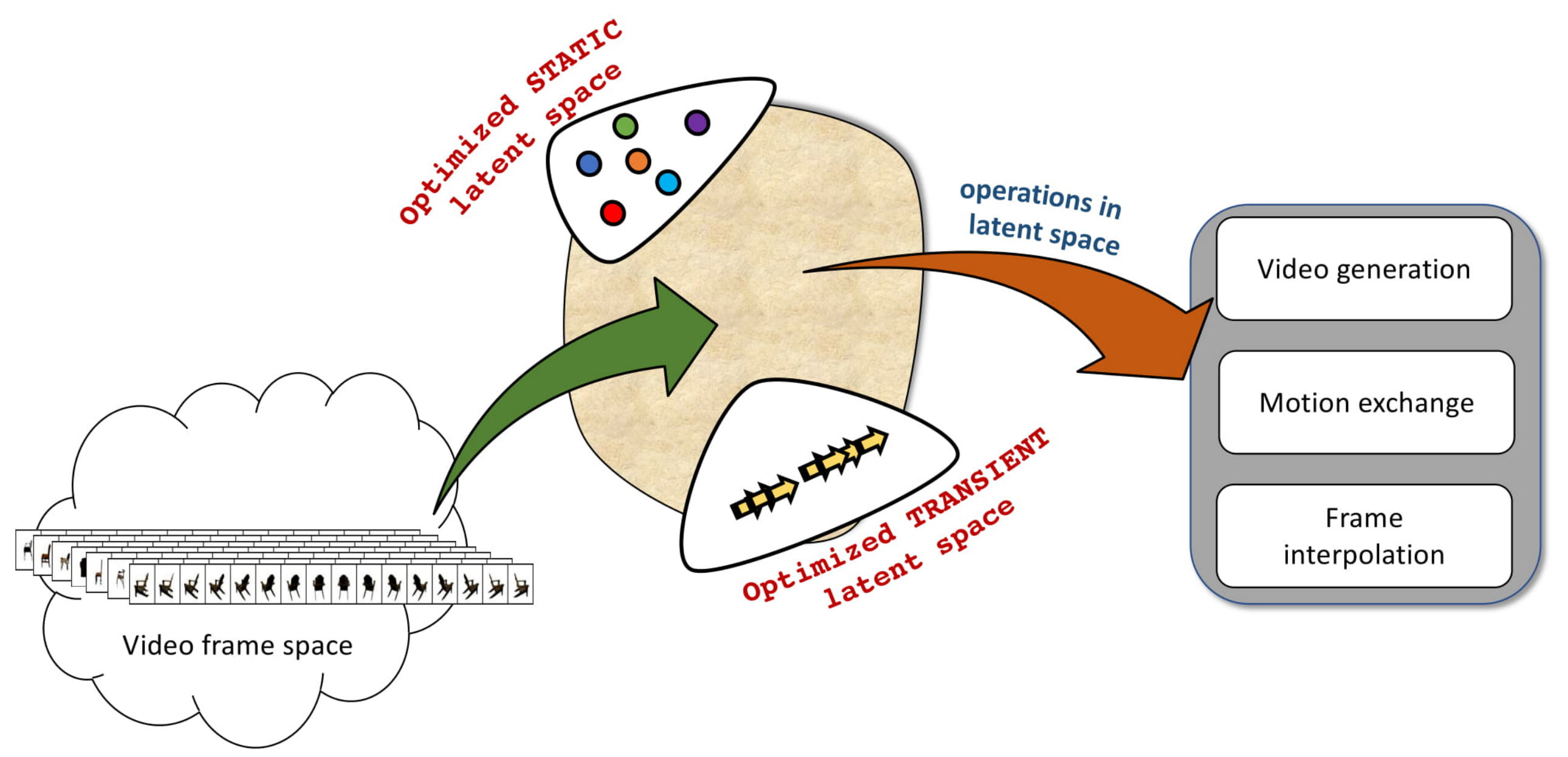
## Why is the Problem Important?

- Understanding how videos are synthesized can help in understanding their spatio-temporal features.
- Can be used as building block for choosing & designing priors.
- Will help us in more difficult problems like future frame prediction, feature learning for videos, etc.

## Related Works

| Methods | Settings | | |
|---|---|---|---|
| | Adversarial learning? | Input frame? | Input latent vectors? |
| VGAN [1] | ✓ | ✗ | ✓(random) |
| TGAN [2] | ✓ | ✗ | ✓(random) |
| MoCoGAN [3] | ✓ | ✗ | ✓(random) |
| Video-VAE [6] | ✗ | ✓ | ✓(random) |
| **Ours** | ✗ | ✗ | ✓(learned) |

## Pictorial Representation


Optimized STATIC latent space
operations in latent space
Video generation
Motion exchange
Frame interpolation
Optimized TRANSIENT latent space
Video frame space

## Learning Loss I

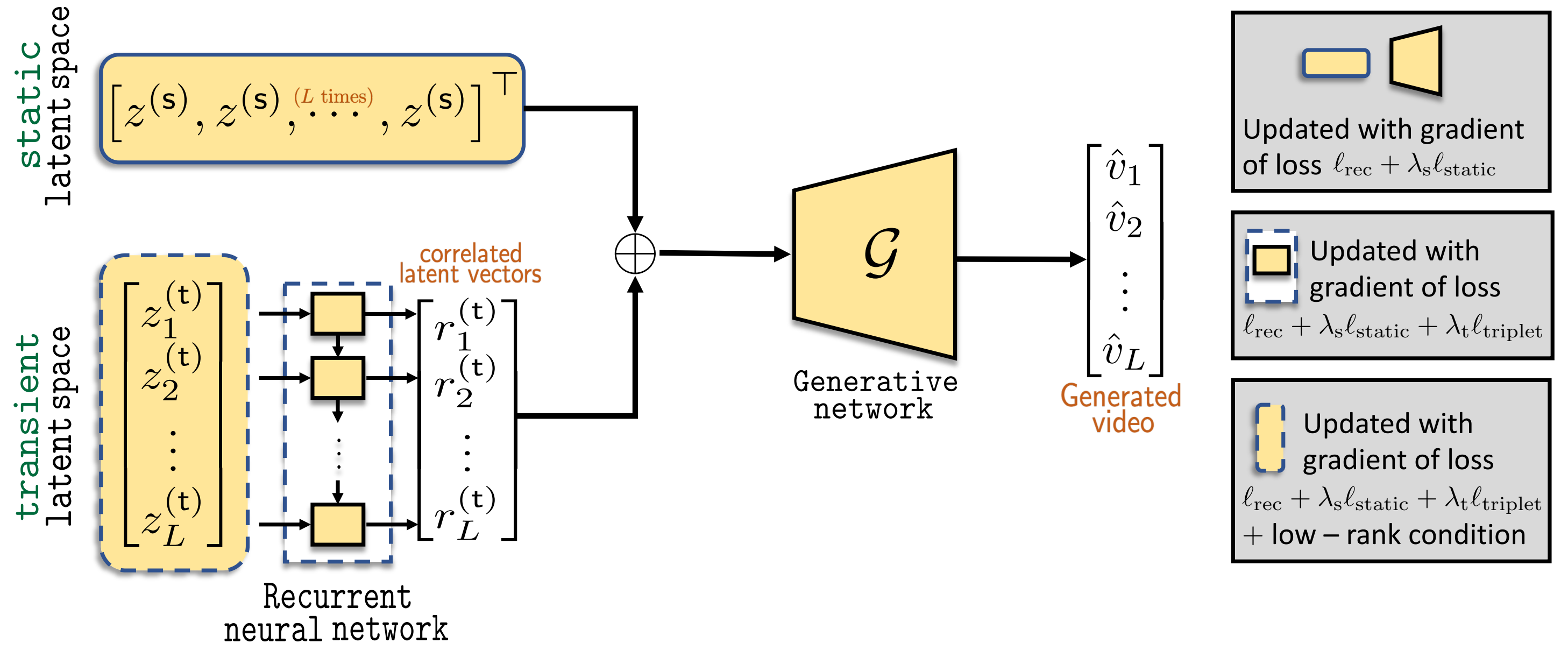
Let weights of $\mathcal{G}$, and the RNN be $\gamma$ and $\theta$, respectively. We jointly optimize for $\theta, \gamma$, and $\{\mathcal{Z}_{\mathsf{V}_j}\}_{j=1}^{N}$ (sampled once in the beginning) in two stages:

$$\text{Stage 1}: \quad \min_{\gamma} \ \ell(\mathsf{V}_j, \mathcal{G}(\mathcal{Z}_{\mathsf{V}_j})|(\mathcal{Z}_{\mathsf{V}_j}, \theta)) \tag{1.1}$$

$$\text{Stage 2}: \quad \min_{\mathcal{Z}_{\mathsf{V}},\theta} \ \ell(\mathsf{V}_j, \mathcal{G}(\mathcal{Z}_{\mathsf{V}_j})|\gamma) \tag{1.2}$$

Index $j$ represents a random video out of $N$ training set videos. $\ell(\cdot)$ can be chosen to be any regression-based loss. We will refer to both (1.1) and (1.2) together as $\min_{\mathcal{Z}_{\mathsf{V}},\theta,\gamma} \ell_{\text{rec}}$.

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. For this, we update the above loss as $\min_{\mathcal{Z}_{\mathsf{V}},\theta,\gamma} \ (\ell_{\text{rec}} + \lambda_{\text{s}}\ell_{\text{static}})$ 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 $\lambda_{\text{s}}$ is the regularization constant.

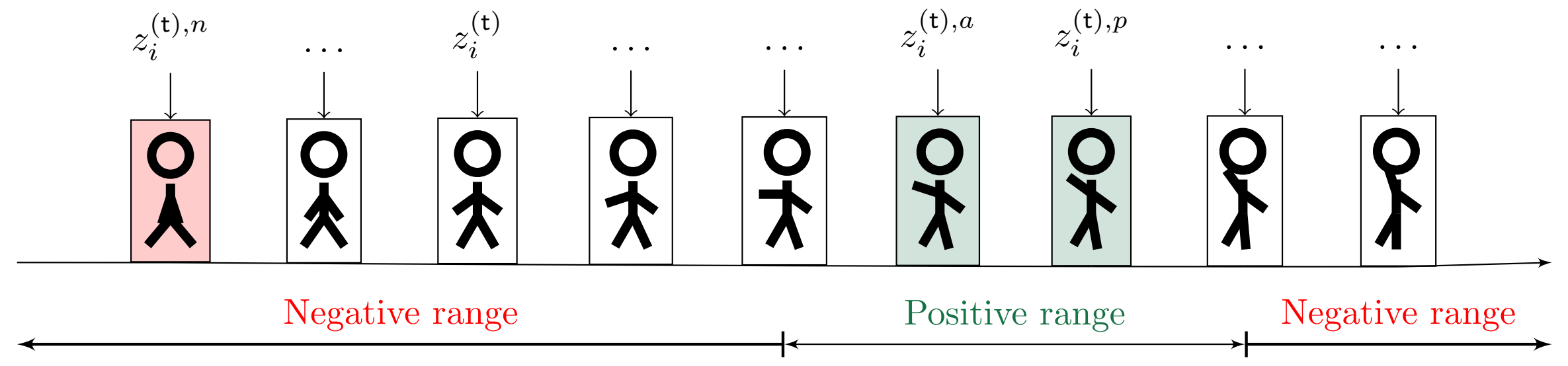

static latent space
$[z^{(\mathsf{s})}, z^{(\mathsf{s})}, \overset{(L \text{ times})}{\cdots}, z^{(\mathsf{s})}]^{\top}$
transient latent space
$z_1^{(\mathsf{t})}$
$z_2^{(\mathsf{t})}$
$z_L^{(\mathsf{t})}$
correlated latent vectors
$r_1^{(\mathsf{t})}$
$r_2^{(\mathsf{t})}$
$r_L^{(\mathsf{t})}$
Recurrent neural network
$\mathcal{G}$
Generative network
$\hat{v}_1$
$\hat{v}_2$
$\hat{v}_L$
Generated video
Updated with gradient of loss $\ell_{\text{rec}} + \lambda_{\text{s}}\ell_{\text{static}}$
Updated with gradient of loss $\ell_{\text{rec}} + \lambda_{\text{s}}\ell_{\text{static}} + \lambda_{\text{t}}\ell_{\text{triplet}}$
Updated with gradient of loss $\ell_{\text{rec}} + \lambda_{\text{s}}\ell_{\text{static}} + \lambda_{\text{t}}\ell_{\text{triplet}}$ + low – rank condition

## Learning Loss II (Triplet Condition)

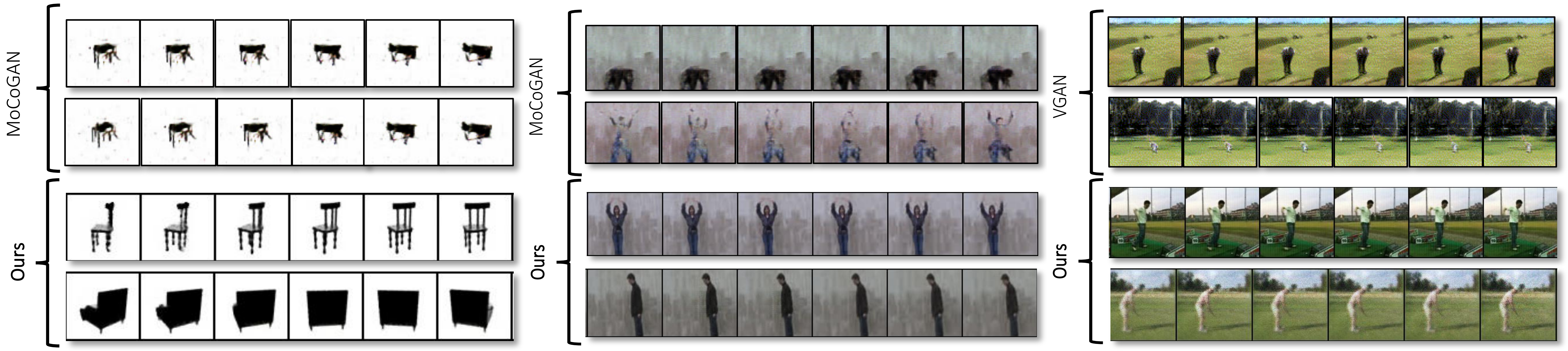
Positive frames are randomly sampled within a margin range $\alpha$ of the anchor and negatives are chosen outside of this margin range. Defining a triplet set with transient latent code vectors $\{z_i^{(\mathsf{t}),a}, z_i^{(\mathsf{t}),p}, z_i^{(\mathsf{t}),n}\}$, we aim to learn the transient embedding space $\mathbf{z}^{(\mathsf{t})}$ such that $\|z_i^{(\mathsf{t}),a} - z_i^{(\mathsf{t}),p}\|_2^2 + \alpha < \|z_i^{(\mathsf{t}),a} - z_i^{(\mathsf{t}),n}\|_2^2 \ \forall \ \{z_i^{(\mathsf{t}),a}, z_i^{(\mathsf{t}),p}, z_i^{(\mathsf{t}),n}\} \in \Gamma$, where $\Gamma$ is the set of all possible triplets in $\mathbf{z}^{(\mathsf{t})}$. With the above regularization and defining $\alpha$ as a hyperparameter that controls the margin while selecting positives and negatives, the above loss can be written as

### Proposed Learning Loss

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



## Qualitative Results



(a) Chair-CAD [7] (b) Weizmann Human Action [8] (c) Golf [1]

## Action exchange

| Actions | Identities | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 |
| run | ● | ● | ● | | 🔴 | | | | |
| walk | 🟢 | ● | | | ● | | ● | | |
| jump | | ● | | ● | ● | ● | 🔵 | ● | ● |
| skip | ● | ● | ● | | | ● | ● | 🟡 | ● |

## Quantitative Results

| | MCS ↓ | FCS ↑ |
|---|---|---|
| Bound | 0.0 | 0.91 |
| MoCo | 4.11 | 0.85 |
| Ours | **3.32** | **0.89** |

(a) Chair-CAD [7]

| | MCS ↓ | FCS ↑ |
|---|---|---|
| Bound | 0.0 | 0.95 |
| MoCo | 3.41 | 0.85 |
| Ours | **2.63** | **0.90** |

(b) Weizmann Human Action [8]

| | MCS ↓ | FCS ↑ |
|---|---|---|
| Bound | 0.0 | 0.97 |
| VGAN | 3.61 | **0.88** |
| Ours | **2.71** | 0.84 |

(c) Golf [1]

## Conclusions

- We presented a non-adversarial approach for synthesizing videos by jointly optimizing both network weights and input latent space.
- Our approach allows to generate videos spanning the diversity of the data distribution, perform frame interpolation, & generate videos unseen during training.

## Acknowledgements

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

## References

[1] Carl Vondrick et al. “Generating videos with scene dynamics”. *NIPS*. 2016.

[2] Masaki Saito et al. “Temporal generative adversarial nets with singular value clipping”. *IEEE ICCV*. 2017.

[3] Sergey Tulyakov et al. “Mocogan: Decomposing motion and content for video generation”. *IEEE CVPR*. 2018.

[4] Piotr Bojanowski et al. “Optimizing the Latent Space of Generative Networks”. *ICML*. 2018.

[5] Yedid Hoshen et al. “Non-Adversarial Image Synthesis with Generative Latent Nearest Neighbors”. *IEEE CVPR*. 2019.

[6] Jiawei He et al. “Probabilistic video generation using holistic attribute control”. *ECCV*. 2018.

[7] Mathieu Aubry et al. “Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models”. *IEEE CVPR*. 2014.

[8] Lena Gorelick et al. “Actions as space-time shapes”. *IEEE TPAMI*. 2007.