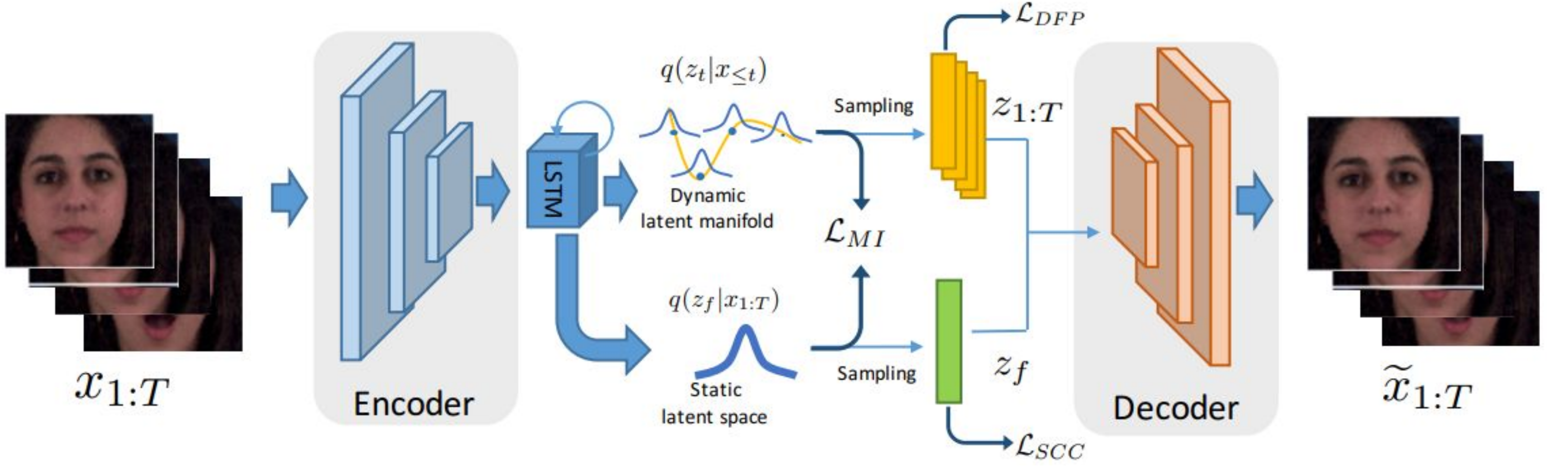
S3VAE: Self-Supervised Sequential VAE for Representation Disentanglement and Data Generation

Yizhe Zhu[1,2], Martin Renqiang Min[1], Asim Kadav[1], Hans Peter Graf[1]
yizhe.zhu@rutgers.edu, {renqiang, asim, hpg}@nec-labs.com
[1]NEC Labs America, [2]Department of Computer Science, Rutgers University

Disentangled Representation Learning: Framework



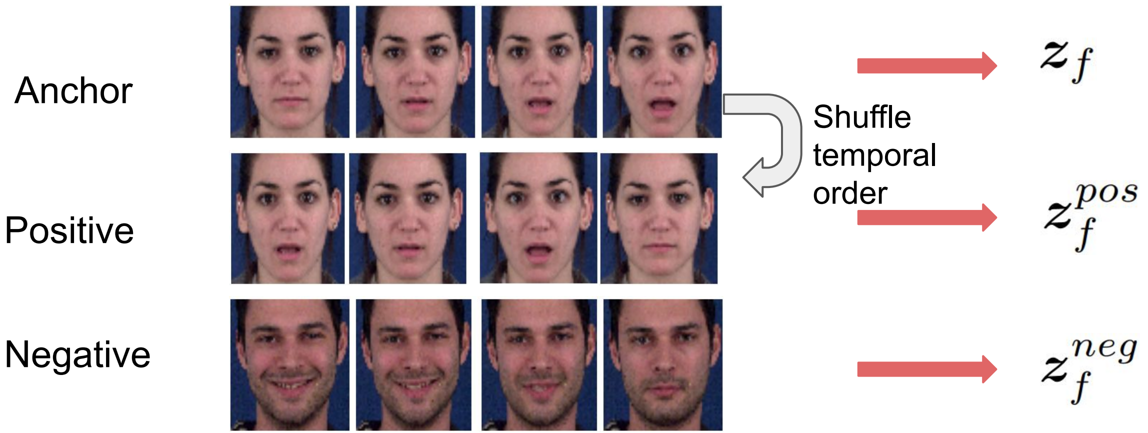
- Encoder
- Decoder
- LSTM in the latent space

VAE Objectives: $$\mathcal{L}_{VAE} = \mathbb{E}_{q(\boldsymbol{z}_{1:T}, \boldsymbol{z}_f | \boldsymbol{x}_{1:T})}[-\sum_{t=1}^{T} \log p(\boldsymbol{x}_t | \boldsymbol{z}_f, \boldsymbol{z}_t)] + \mathrm{KL}(q(\boldsymbol{z}_f | \boldsymbol{x}_{1:T}) || p(\boldsymbol{z}_f)) + \sum_{t=1}^{T} \mathrm{KL}(q(\boldsymbol{z}_t | \boldsymbol{x}_{\leq t}) || p(\boldsymbol{z}_t | \boldsymbol{z}_{<t}))$$

Self-Supervised Signal (1): Static Consistency Constraint

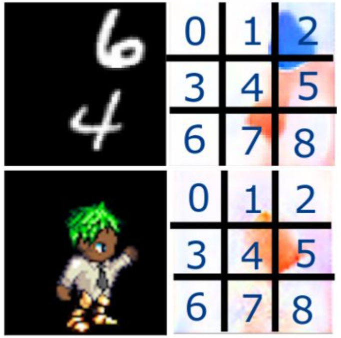
- To encourage the appearance representation $\boldsymbol{z}_f$ to exclude any dynamic information.
- Triplet Loss:

$$\mathcal{L}_{SCC} = \max\left(D(\boldsymbol{z}_f, \boldsymbol{z}_f^{pos}) - D(\boldsymbol{z}_f, \boldsymbol{z}_f^{neg}) + \boldsymbol{m}, 0\right)$$

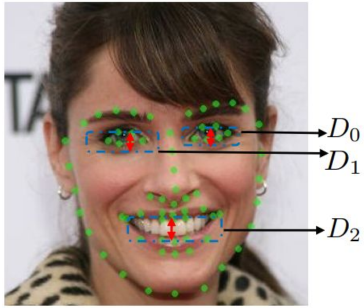


Self-Supervised Signal (2): Dynamic Factor Prediction

- To encourage the motion representation z_t to carry adequate and correct time-dependent information of each timestep
- **Optical flow** provides the location of motion
  - Grid the optical flow map with indices
- **Landmarks** provides the subtle motion on facial expression
  - Distances between upper and lower eyelips and distances between lips



The input frame and optical flow



The three distances on faces

Self-Supervised Signal (3): Mutual Information

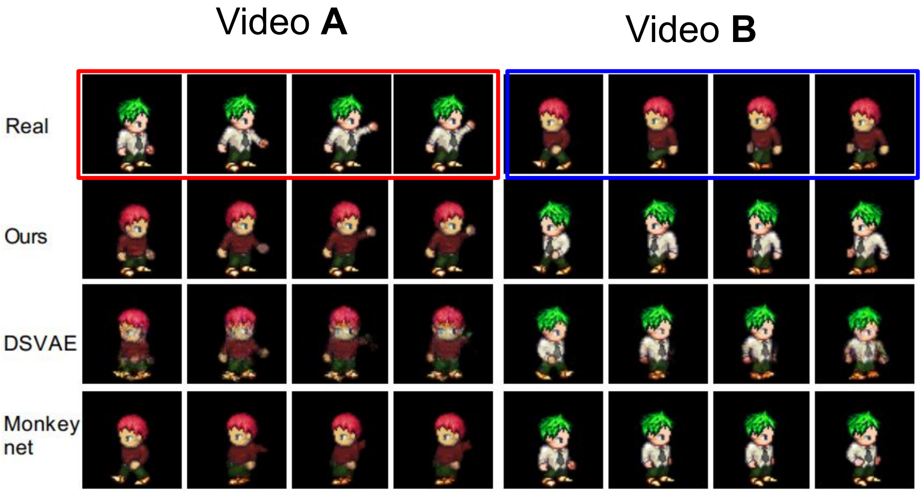
- To encourage the information in z_f and z_t to be mutually exclusive.
- To minimize the mutual information between z_f and z_t

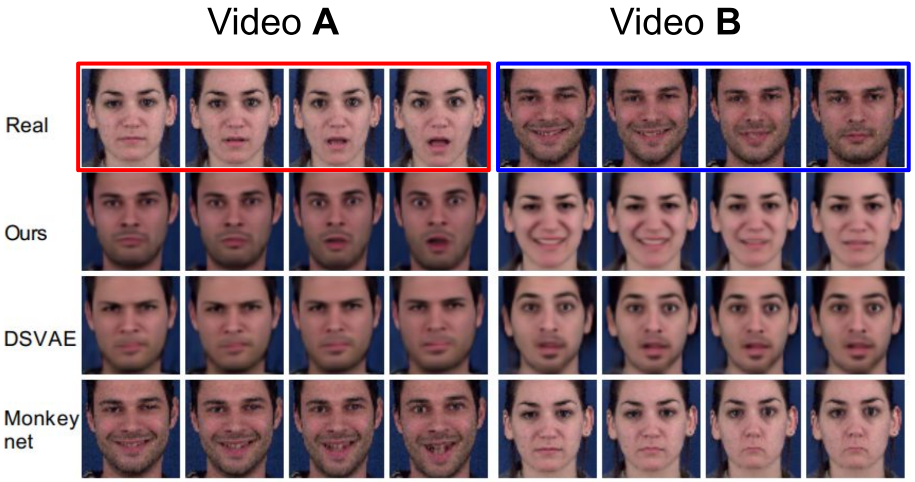
$$\begin{aligned}\mathcal{L}_{MI}(\boldsymbol{z}_f, \boldsymbol{z}_{1:T}) &= \sum_{t=1}^{T} \mathrm{KL}(q(\boldsymbol{z}_f, \boldsymbol{z}_t) || q(\boldsymbol{z}_f) q(\boldsymbol{z}_t)) \\ &= \sum_{t=1}^{T} [f(q(\boldsymbol{z}_f, \boldsymbol{z}_t)) - f(q(\boldsymbol{z}_f)) - f(q(\boldsymbol{z}_t))],\end{aligned}$$

$$\text{where } f(q(\cdot)) = \mathbb{E}_{q(\boldsymbol{z})}[\log(\cdot)] = \mathbb{E}_{q(\boldsymbol{z}_f, \boldsymbol{z}_t)}[\log(\cdot)].$$

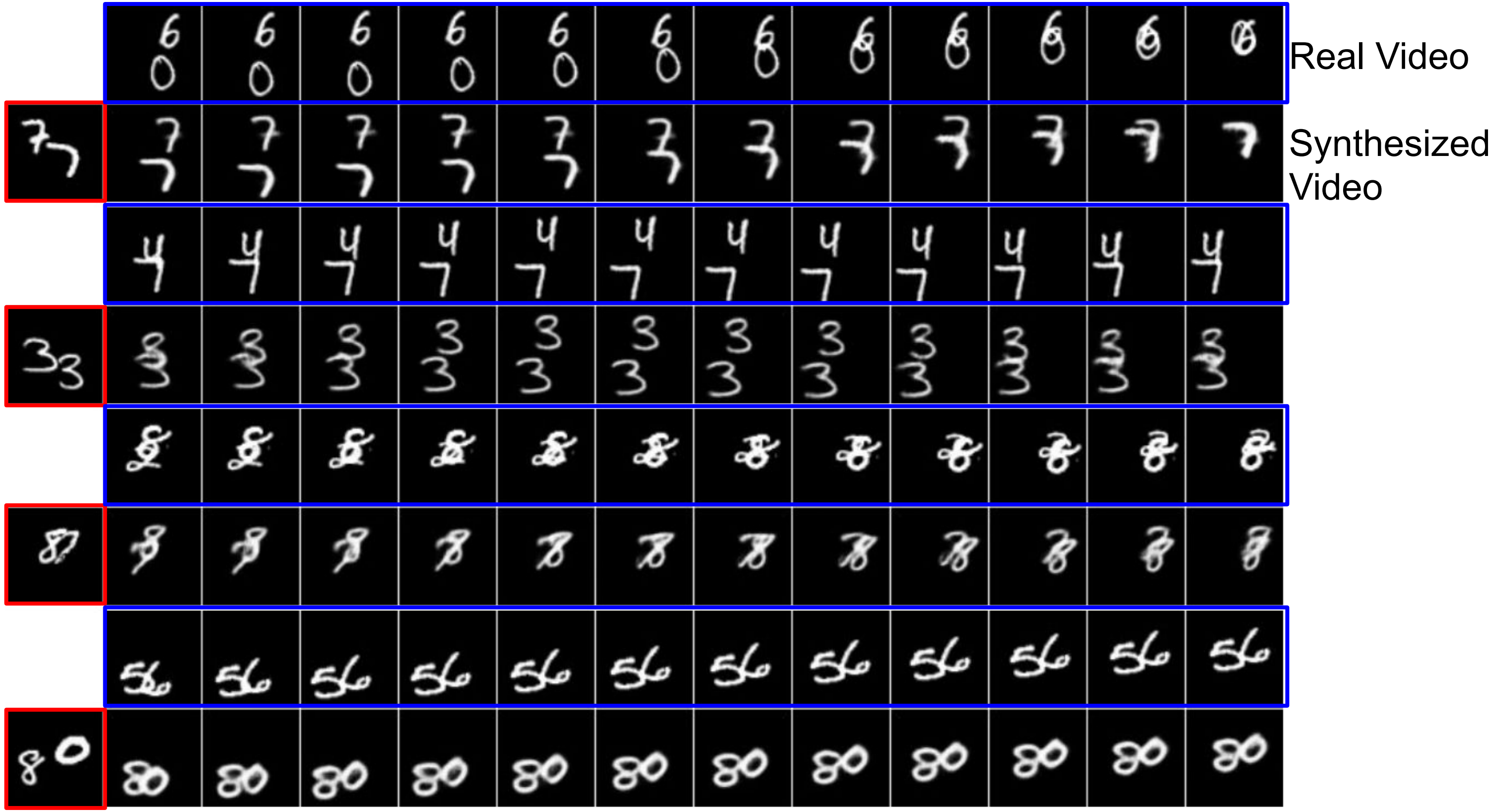
Experiments: Representation Swapping

- Swap the appearance and motion representation of two given videos





Experiments: Representation Swapping

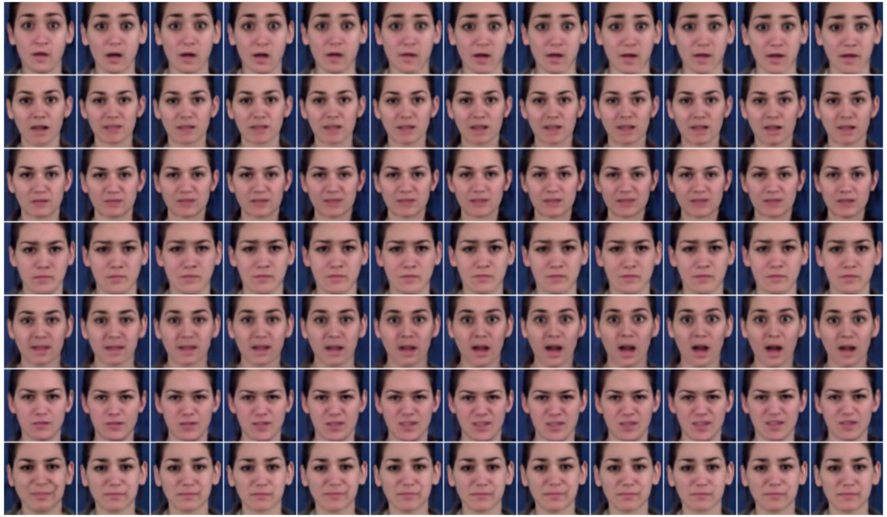


Experiments: Manipulating video generation(Dsprite)

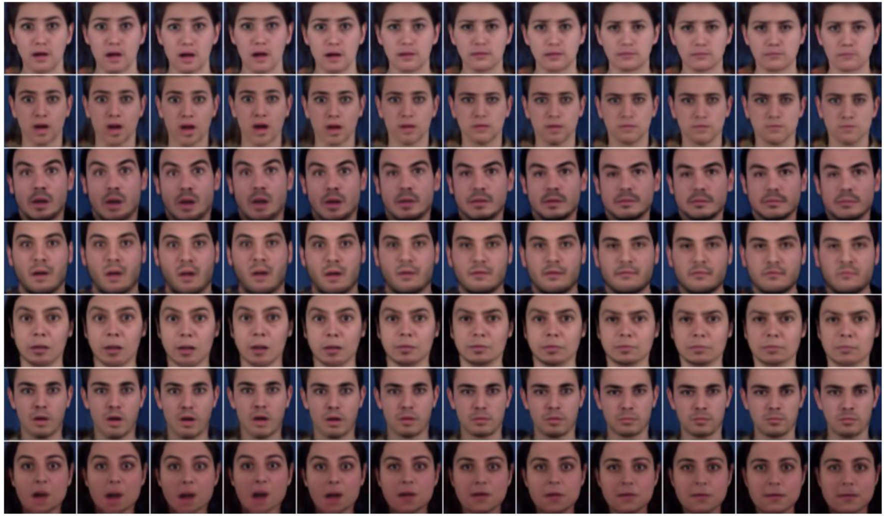
Fix appearance representation

Fix motion representation

Experiments: Manipulating video generation (MUG)



Fix appearance representation



Fix motion representation

Experiments: Quantitatively performance comparison

Table 1. Quantitatively performance comparison on SMMNIST, Sprite and MUG datasets. High values are expected for Acc, $H(y)$ and IS, while for $H(y|x)$, the lower values are better. The results of our model with supervision of ground truth labels *baseline-sv** are shown as a reference.

Methods	SMMNIST				Sprite				MUG			
	Acc	IS	$H(y\|x)$	$H(y)$	Acc	IS	$H(y\|x)$	$H(y)$	Acc	IS	$H(y\|x)$	$H(y)$
MoCoGAN	74.55%	4.078	0.194	0.191	92.89%	8.461	0.090	2.192	63.12%	4.332	0.183	1.721
DSVAE	88.19%	6.210	0.185	2.011	90.73%	8.384	0.072	2.192	54.29%	3.608	0.374	1.657
baseline	90.12%	6.543	0.167	2.052	91.42%	8.312	0.071	2.190	53.83%	3.736	0.347	1.717
full model	**95.09%**	**7.072**	**0.150**	**2.106**	**99.49%**	**8.637**	**0.041**	**2.197**	**70.51%**	**5.136**	**0.135**	**1.760**
*baseline-sv**	92.18%	6.845	0.156	2.057	98.91%	8.741	0.028	2.196	72.32%	5.006	0.129	1.740

- *Baseline:* our sequential VAE without self-supervision
- *Baseline-sv:* our sequential VAE with supervision of ground truth labels
- *Full model:* our sequential VAE with self-supervision