RSGAN: Recurrent Stacked Generative Adversarial Network for Conditional Video Generation

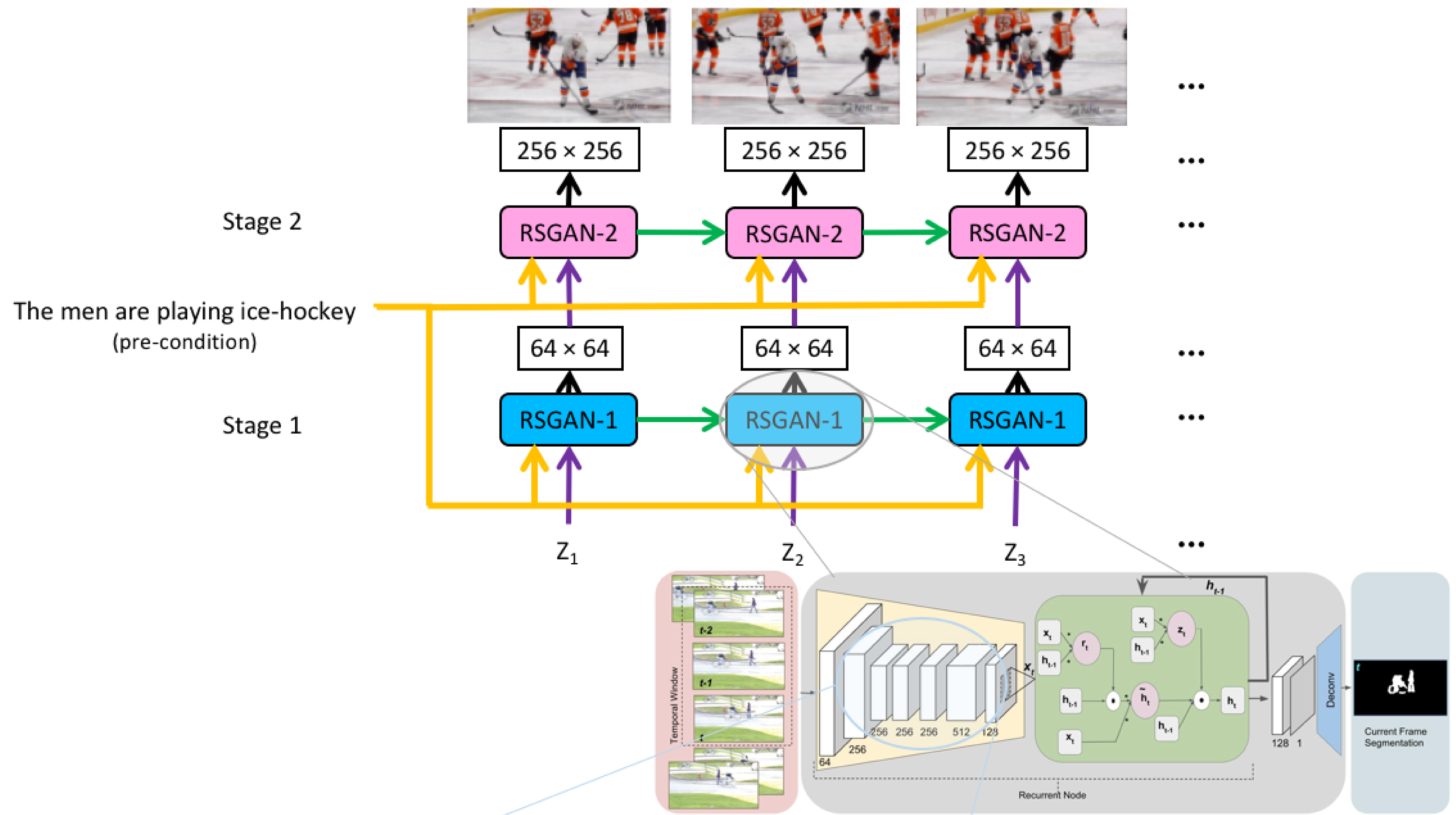
S. Naha, K. M. Nayem and M. L. Islam   School of Informatics and Computing, Indiana University

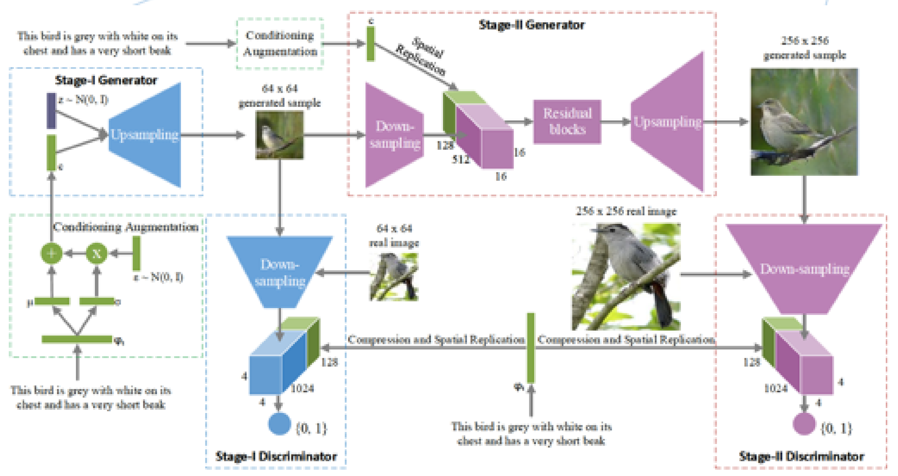
Introduction

Generating video frames based on a pre-condition is a challenging problem and requires understanding of per frame contents and visual dynamics and their relevacies to the pre-condition. In this project, we propose a novel Recurrent Stacked Generative Adversarial Network (RSGAN) based model to generate video frames based on a given pre-condition. In our knowledge, this is the first work to address the problem of conditional video generation using adversarial network. We can address the problem of generating videos based on pre-conditions such as,

1. action classes
2. fMRI signals
3. sentence descriptions

Our Approach





Each RSGAN module is - consist of a StackGAN model.
- connected by a Fully Convolutional LSTM Network.

Objective Function

Conditioned on Gaussian latent variables c_0, Stage-I RSGAN trains discriminator D_0 and generator G_0 by alternatively maximizing $\mathcal{L}_{D_o}$ and minimizing $\mathcal{L}_{G_o}$.

Stage-I RSGAN:

$$\mathcal{L}_{D_o} = \mathbb{E}_{(I_o,t)\sim p_{data}}[logD_0(I_0,\varphi_t)] + \mathbb{E}_{z\sim p_z, t\sim p_{data}}[log(1-D_0(G_0(z,c_0),\varphi_t))]$$

$$\mathcal{L}_{G_o} = \mathbb{E}_{z\sim p_z, t\sim p_{data}}[log(1-D_0(G_0(z,c_0),\varphi_t))] + \lambda D_{KL}(\mathcal{N}(\mu_0(\varphi_t),\textstyle\sum_0(\varphi_t))||\mathcal{N}(0,I))$$

Conditioned on the low resulation sample s_0 and Gaussian latent variables c, discriminator D and generator G in Stage-II RSGAN is trained by alternatively maximizing $\mathcal{L}_D$ and minimizing $\mathcal{L}_G$.

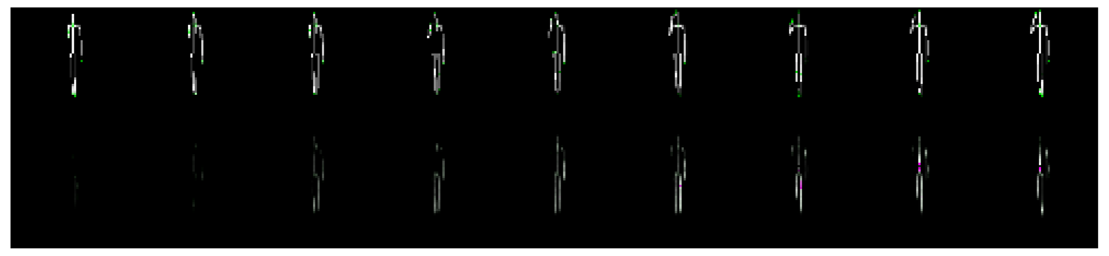
Stage-II RSGAN:

$$\mathcal{L}_D = \mathbb{E}_{(I,t)\sim p_{data}}[logD(I,\varphi_t)] + \mathbb{E}_{s_o\sim pG_o, t\sim p_{data}}[log(1-D(G(s_0,c),\varphi_t))]$$

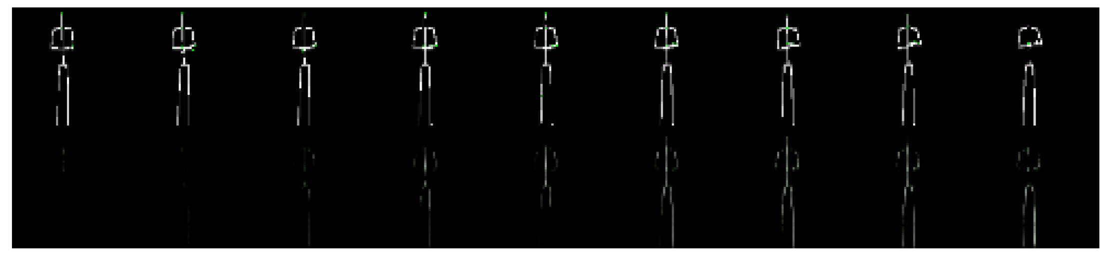
$$\mathcal{L}_G = \mathbb{E}_{s_o\sim pG_o, t\sim p_{data}}[log(1-D(G(s_0,c),\varphi_t))] + \lambda D_{KL}(\mathcal{N}(\mu(\varphi_t),\textstyle\sum(\varphi_t))||\mathcal{N}(0,I))$$

Result

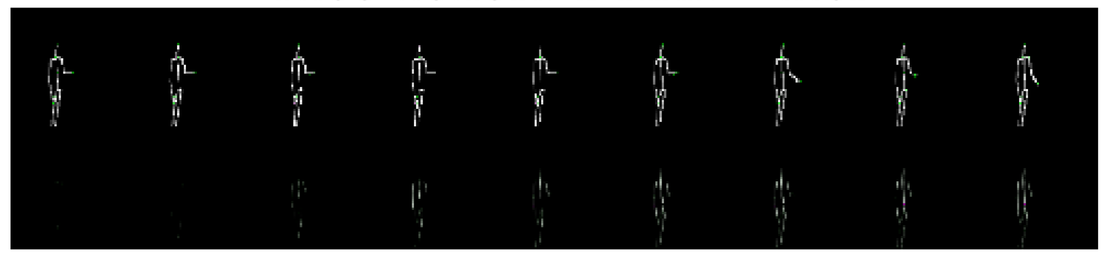
Ground truth (1^{st} row) and the **partial result on Convolutional LSTM** (2^{nd} row).



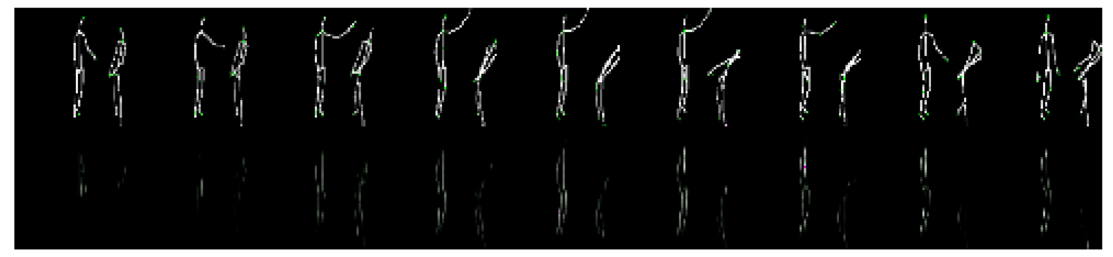
Video action: Drink water



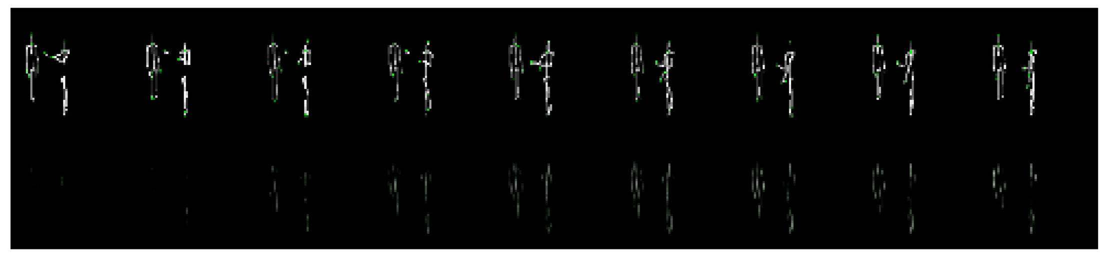
Video action: Reading



Video action: Drop



Video action: Slapping other person



Video action: Point finger at the other person

Video action: Writing

- Right now, we are trying to generate video with simple details. That's why we are using NTU RGB+D Action Recognition Dataset (skeletal data).

Future Research

- Generate video with complex details and multiple moving objects.
- Use fMRI dataset of human brain, to generate video.

References

[1] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. *arXiv preprint arXiv:1612.03242*, 2016.

[2] S. Valipour, M. Siam, M. Jagersand, and N. Ray. Recurrent Fully Convolutional Networks for Video Segmentation. *arXiv preprint arXiv:1611.09904*, 2016.

[3] O. Mogren. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904*, 2016.