Deep Learning for Computer Vision

Deep Generative Models: Video Applications

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad

Generating Videos with Scene Dynamics[1]

Can we use GANs to generate videos?

[1] Vondrick et al, Generating Videos with Scene Dynamics, NeurIPS 2016

Generating Videos with Scene Dynamics[1]

Can we use GANs to generate videos?

Recall GAN objective (w.r.t corresponding G and D parameters):

$$\min_{w_G} \max_{w_D} \mathbb{E}_{x \sim p_x(x)}[\log D(x; w_D)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z; w_G); w_D))]$$

G and D can take on any form appropriate for a task as long as they are differentiable w.r.t. parameters w_G and w_D

[1] Vondrick et al, Generating Videos with Scene Dynamics, NeurIPS 2016

Generating Videos with Scene Dynamics[1]

Can we use GANs to generate videos?

Recall GAN objective (w.r.t corresponding G and D parameters):

$$\min_{w_G} \max_{w_D} \mathbb{E}_{x \sim p_x(x)}[\log D(x; w_D)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z; w_G); w_D))]$$

G and D can take on any form appropriate for a task as long as they are differentiable w.r.t. parameters w_G and w_D

Consider the output of G to be:

$$G(z) = m(z) \odot f(z) + (1 - m(z)) \odot b(z)$$

where f is **Foreground**, b is **Background**, and m is a **Mask** which indicates whether to use foreground or background for a pixel

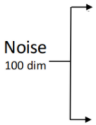
[1] Vondrick et al, Generating Videos with Scene Dynamics, NeurIPS 2016

Generating Videos with Scene Dynamics: Generator[2]

Noise
100 dim

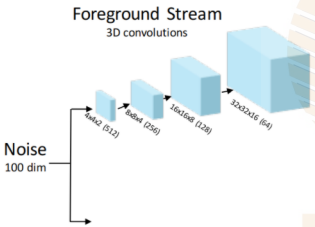
$$m \odot f + (1 - m) \odot b$$

Generating Videos with Scene Dynamics: Generator[2]



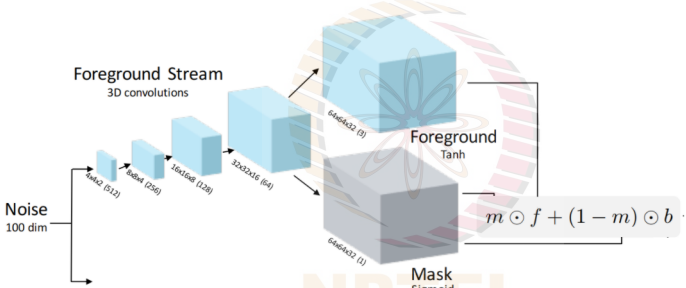
$$m \odot f + (1 - m) \odot b$$

Generating Videos with Scene Dynamics: Generator[2]

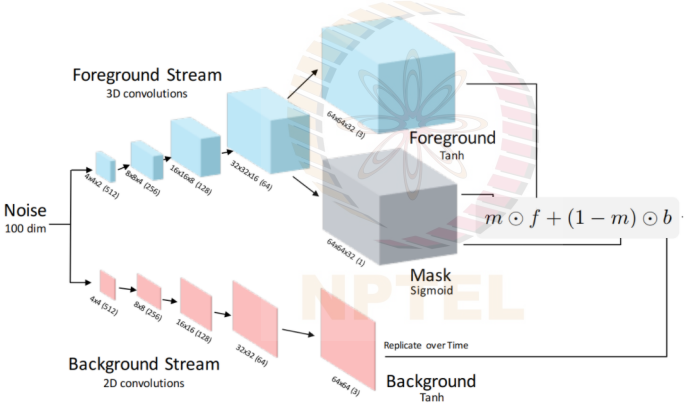


$$m \odot f + (1 - m) \odot b$$

Generating Videos with Scene Dynamics: Generator[2]

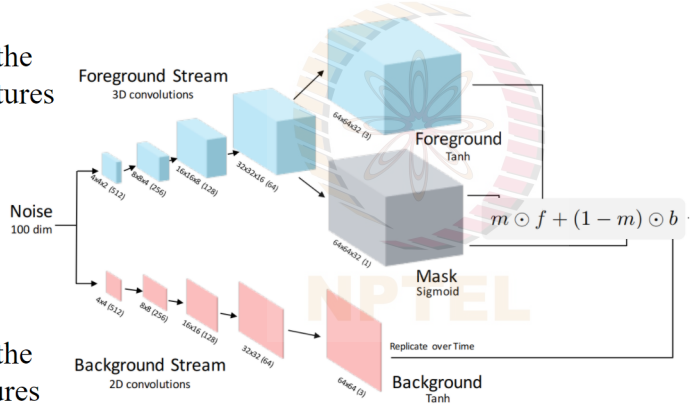


Generating Videos with Scene Dynamics: Generator[2]



Generating Videos with Scene Dynamics: Generator[2]

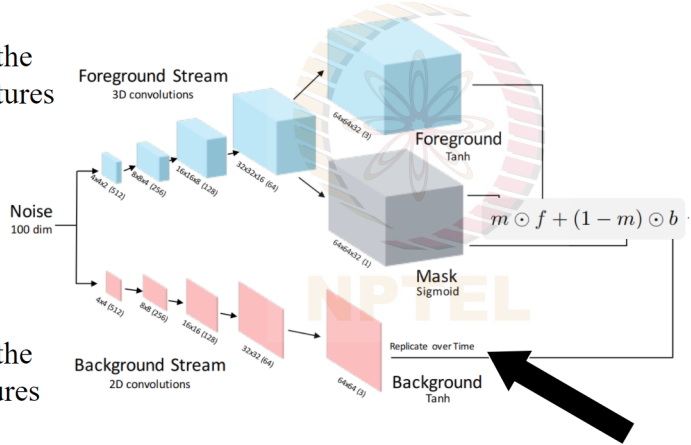
Captures the moving features



Captures the static features

Generating Videos with Scene Dynamics: Generator[2]

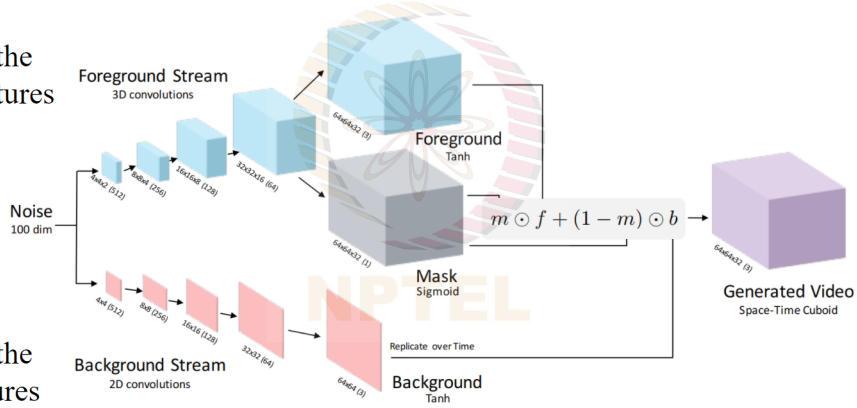
Captures the moving features



Captures the static features

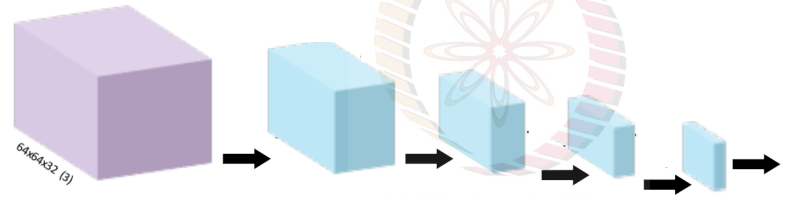
Generating Videos with Scene Dynamics: Generator[2]

Captures the moving features

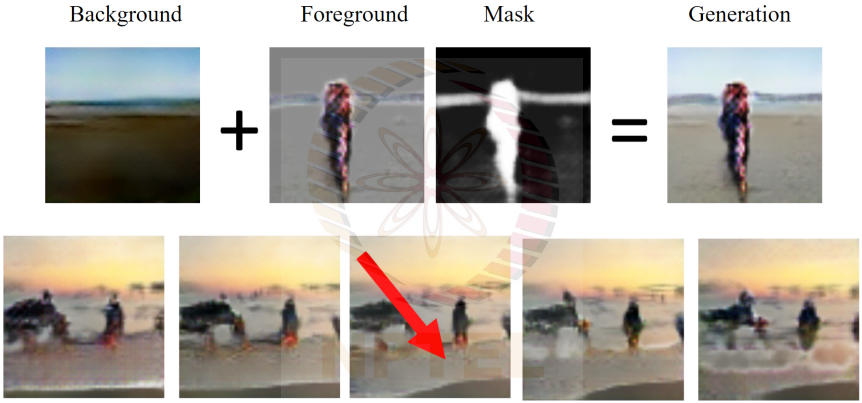


Captures the static features

Generating Videos with Scene Dynamics: Discriminator



Generating Videos with Scene Dynamics: Results[3]

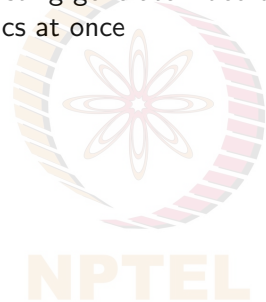


For more examples, see http://www.cs.columbia.edu/~vondrick/tinyvideo/

[3]Vondrick et al, Generating Videos with Scene Dynamics, NeurIPS 2016

The Pose Knows: Video Forecasting by Generating Pose Futures[4]

- GANs and VAEs in video forecasting generate video directly in pixel space $\implies$ model all the structure and scene dynamics at once



[4]Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

The Pose Knows: Video Forecasting by Generating Pose Futures[4]

- GANs and VAEs in video forecasting generate video directly in pixel space $\implies$ model all the structure and scene dynamics at once
- In unconstrained settings, often generate uninterpretable results

[4]Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

The Pose Knows: Video Forecasting by Generating Pose Futures[4]

- GANs and VAEs in video forecasting generate video directly in pixel space $\implies$ model all the structure and scene dynamics at once
- In unconstrained settings, often generate uninterpretable results

Solution

- Forecasting needs to be done first at a higher level of abstraction (pose)

[4] Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

The Pose Knows: Video Forecasting by Generating Pose Futures[4]

- GANs and VAEs in video forecasting generate video directly in pixel space $\implies$ model all the structure and scene dynamics at once
- In unconstrained settings, often generate uninterpretable results

Solution

- Forecasting needs to be done first at a higher level of abstraction (pose)
- Exploit human pose detectors as (free) source of supervision, and break video forecasting problem into two steps:

[4] Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

The Pose Knows: Video Forecasting by Generating Pose Futures[4]

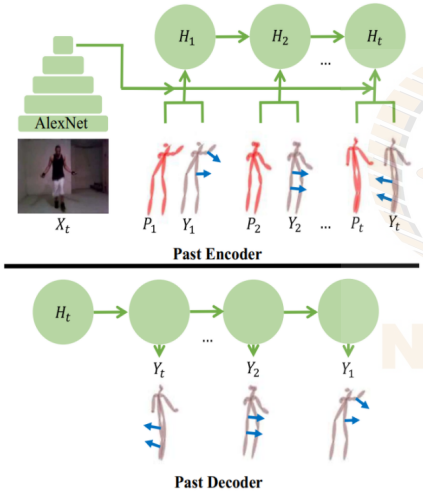
- GANs and VAEs in video forecasting generate video directly in pixel space $\implies$ model all the structure and scene dynamics at once
- In unconstrained settings, often generate uninterpretable results

Solution

- Forecasting needs to be done first at a higher level of abstraction (pose)
- Exploit human pose detectors as (free) source of supervision, and break video forecasting problem into two steps:
  - Use a VAE to model the possible movements of human in pose space
  - Use generated future poses as conditional information to a GAN to predict future frames in pixel space

[4] Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

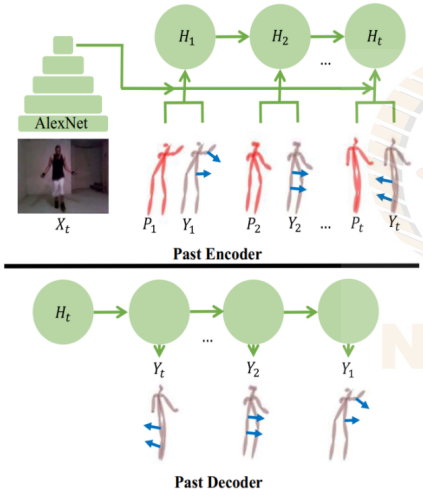
Pose Prediction: Encoder-Decoder Model[5]



- Reads in image features from X_t, corresponding past poses $P_{1..t}$ and their corresponding velocities $Y_{1..t}$

[5] Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

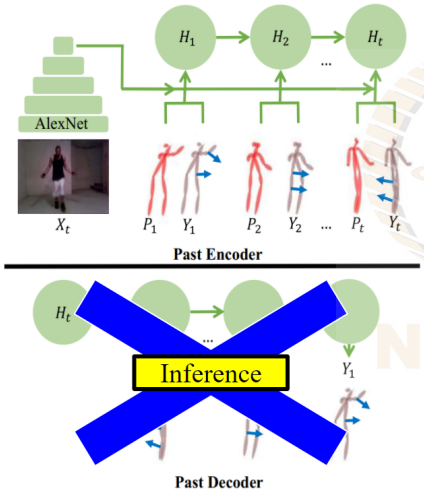
Pose Prediction: Encoder-Decoder Model[5]



- Reads in image features from X_t, corresponding past poses $P_{1..t}$ and their corresponding velocities $Y_{1..t}$
- Replays pose velocities $Y_{1..t}$ in reverse order

[5] Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

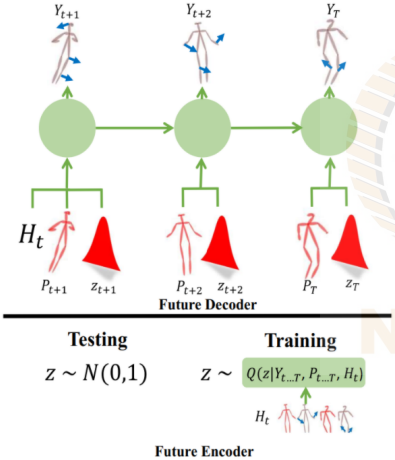
Pose Prediction: Encoder-Decoder Model[5]



- Reads in image features from X_t, corresponding past poses $P_{1..t}$ and their corresponding velocities $Y_{1..t}$
- Replays pose velocities $Y_{1..t}$ in reverse order

[5] Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

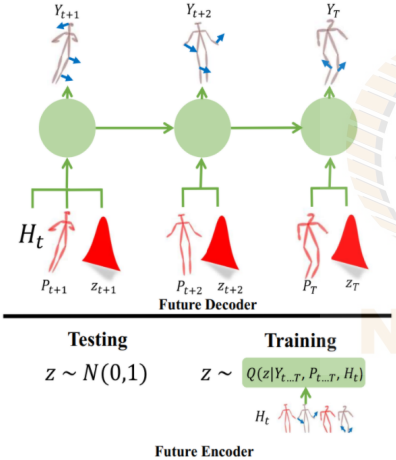
Pose Prediction: Encoder-Decoder Model[6]



- **Inputs:** Past H_t, future pose information $Y_{t+1...T}$, $P_{t+1...T}$; **Output:** Approximate posterior Q

[6] Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

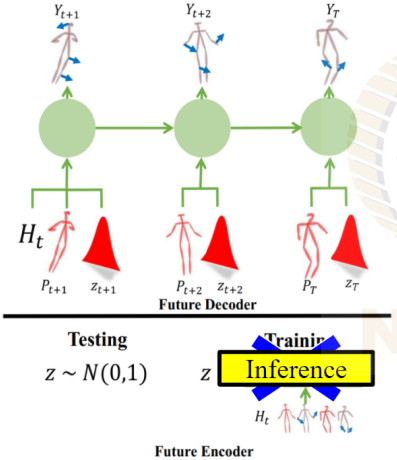
Pose Prediction: Encoder-Decoder Model[6]



- Samples z from Q to reconstruct pose motions $Y_{t+1...T}$ given past H_t and poses $t+1...T$
- **Inputs:** Past H_t, future pose information $Y_{t+1...T}$, $P_{t+1...T}$; **Output:** Approximate posterior Q

[6] Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

Pose Prediction: Encoder-Decoder Model[6]



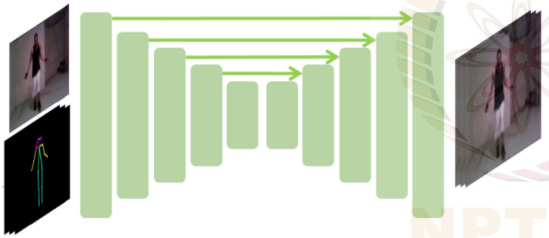
- Samples z from Q to reconstruct pose motions $Y_{t+1...T}$ given past H_t and poses $_{t+1...T}$

- **Inputs:** Past H_t, future pose information $Y_{t+1...T}$, $P_{t+1...T}$; **Output:** Approximate posterior Q

[6] Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

Video Generation[7]

$$L_G = \sum_{i=M/2+1}^{M} l(D(G(I, S_T)), l_r) + \alpha ||G(I, S_T) - V_i||_1$$



$$L_D = \sum_{i=1}^{M/2} l(D(V_i), l_r) + \sum_{i=M/2+1}^{M} l(D(G(I, S_T)), l_f)$$

V: Ground Truth Video

M: Batch size

I: Input

S_T: Pose skeleton

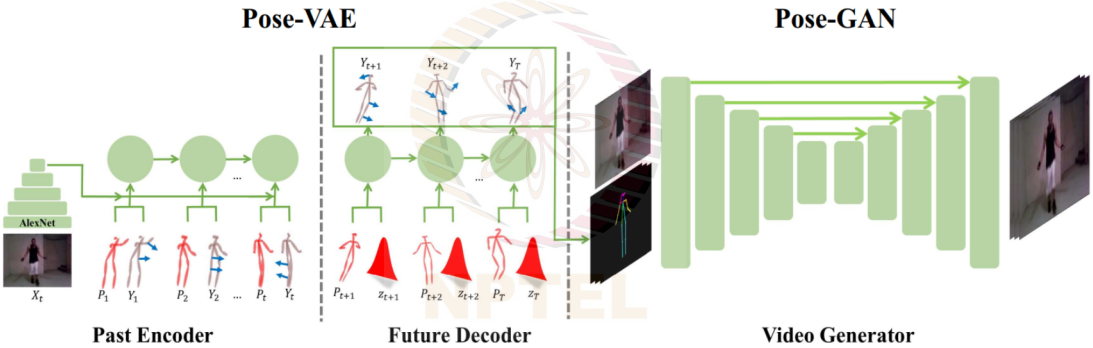
l_r: Real label (1)

l_f: Fake label (0)

L: Binary cross-entropy loss

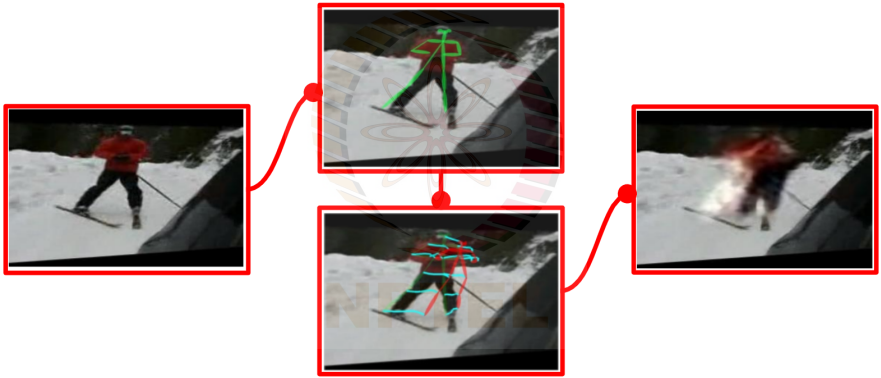
[7] Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

The Pose Knows: Video Forecasting by Generating Pose Futures[8]



[8]Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

Results[9]



[9]Walker et al, The Pose Knows: Video Forecasting by Generating Pose Futures, ICCV 2017

Everybody Dance Now[10]

Objective

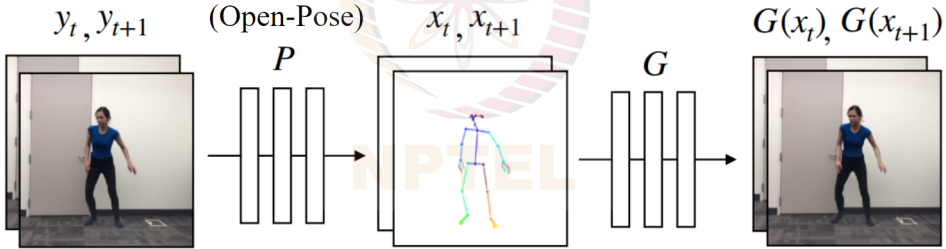
Given a professional's dancing video and an amateur's dancing video, can we generate a video of an amateur dancing professionally?

[10] Chan et al, Everybody Dance Now, ICCV 2019

Everybody Dance Now[10]

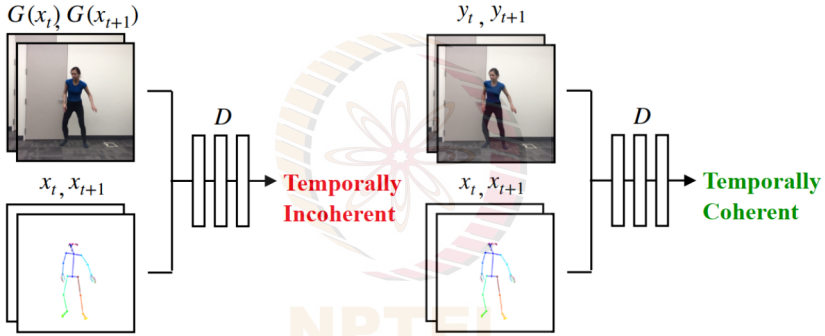
Objective

Given a professional's dancing video and an amateur's dancing video, can we generate a video of an amateur dancing professionally?



[10] Chan et al, Everybody Dance Now, ICCV 2019

Everybody Dance Now: Discriminator[11]



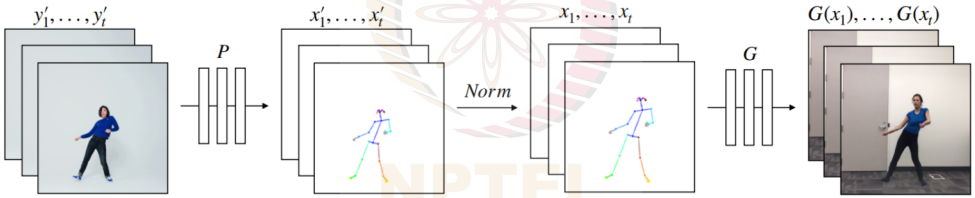
Enforcing temporal coherence between adjacent frames:

$$L_{\text{smooth}}(G, D) = \mathbb{E}_{(x,y)}[\log D(x_t, x_{t+1}, y_t, y_{t+1})] + \mathbb{E}_x[\log(1 - D(x_t, x_{t+1}, G(x_t), G(x_{t+1}))]$$

[11] Chan et al, Everybody Dance Now, ICCV 2019

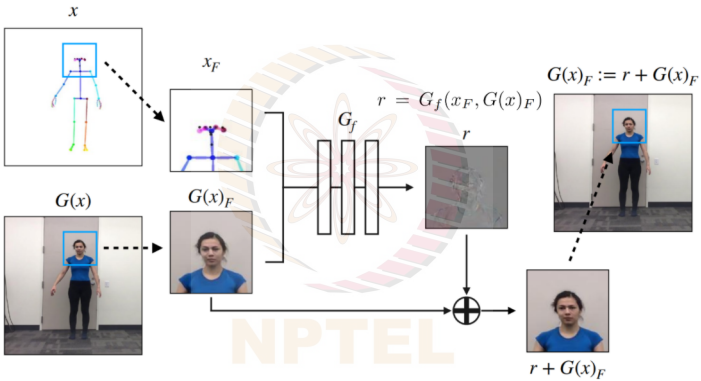
Everybody Dance Now: Inference[12]

Different people may have different limb proportion $\implies$ normalization layer in between



[12] Chan et al, Everybody Dance Now, ICCV 2019

Everybody Dance Now: Refining Generated Face[13]



$$\mathcal{L}_{\text{face}}(G_f, D_f) = \mathbb{E}_{(x_F, y_F)}[\log D_f(x_F, y_F)]$$
$$+\mathbb{E}_{x_F}[\log\left(1 - D_f(x_F, G(x)_F + r)\right)].$$

Everybody Dance Now: Overview

Stage 1

$$\min_G((\max_{D_i}\sum_{k_i}\mathcal{L}_{\text{smooth}}(G,D_k)) + \lambda_{FM}\sum_{k_i}\mathcal{L}_{\text{FM}}(G,D_k)$$
$$+\lambda_P(\mathcal{L}_P(G(x_{t-1}),y_{t-1}) + \mathcal{L}_P(G(x_t),y_t)))$$

Everybody Dance Now: Overview

Stage 1

$$\min_G((\max_{D_i}\sum_{k_i}\mathcal{L}_{\text{smooth}}(G,D_k))+\lambda_{FM}\sum_{k_i}\mathcal{L}_{\text{FM}}(G,D_k)$$
$$+\lambda_P(\mathcal{L}_P(G(x_{t-1}),y_{t-1})+\mathcal{L}_P(G(x_t),y_t)))$$

$$\mathcal{L}_{\text{smooth}}(G,D)=\mathbb{E}_{(x,y)}[\log D(x_t,x_{t+1},y_t,y_{t+1})]$$
$$+\mathbb{E}_x[\log(1-D(x_t,x_{t+1},G(x_t),G(x_{t+1}))]$$

$\mathcal{L}_{\text{FM}}(G,D)$ Discriminator Feature-matching loss (as in Pix2Pix)

$\mathcal{L}_P(G(x),y)$ Perceptual Reconstruction Loss

Everybody Dance Now: Overview

Stage 1

Freeze Stage 1 weights

$$\min_G((\max_{D_i}\sum_{k_i}\mathcal{L}_{\text{smooth}}(G,D_k)) + \lambda_{FM}\sum_{k_i}\mathcal{L}_{\text{FM}}(G,D_k) + \lambda_P(\mathcal{L}_P(G(x_{t-1}),y_{t-1}) + \mathcal{L}_P(G(x_t),y_t)))$$

Everybody Dance Now: Overview

Stage 1

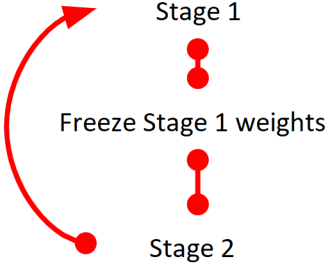
Freeze Stage 1 weights

Stage 2

$$\min_G((\max_{D_i}\sum_{k_i}\mathcal{L}_{\text{smooth}}(G,D_k)) + \lambda_{FM}\sum_{k_i}\mathcal{L}_{\text{FM}}(G,D_k) + \lambda_P(\mathcal{L}_P(G(x_{t-1}),y_{t-1}) + \mathcal{L}_P(G(x_t),y_t)))$$

$$\min_{G_f}\left(\left(\max_{D_f}\mathcal{L}_{\text{face}}(G_f,D_f)\right) + \lambda_P\mathcal{L}_P(r + G(x)_F, y_F)\right)$$

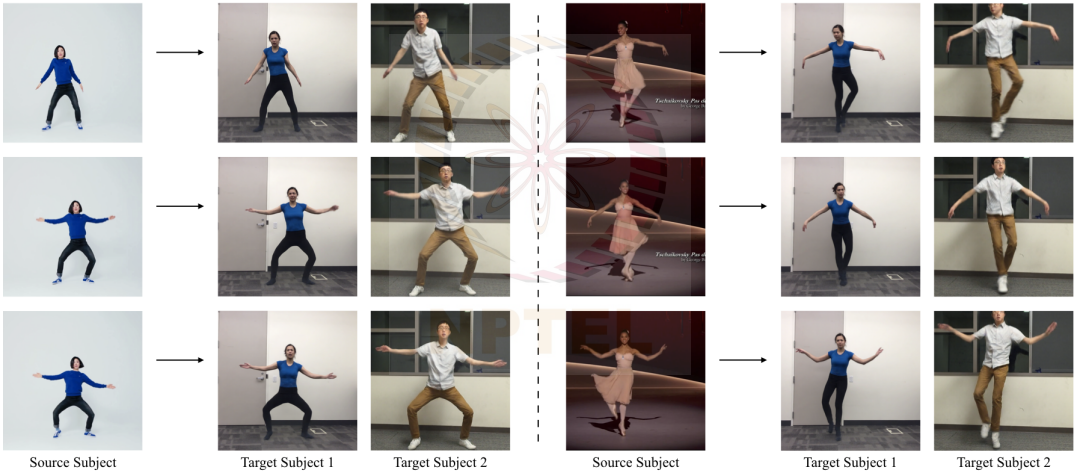
Everybody Dance Now: Overview



$$\min_G((\max_{D_i}\sum_{k_i}\mathcal{L}_{\text{smooth}}(G,D_k)) + \lambda_{FM}\sum_{k_i}\mathcal{L}_{\text{FM}}(G,D_k) + \lambda_P(\mathcal{L}_P(G(x_{t-1}),y_{t-1}) + \mathcal{L}_P(G(x_t),y_t)))$$

$$\min_{G_f}\left(\left(\max_{D_f}\mathcal{L}_{\text{face}}(G_f,D_f)\right) + \lambda_P\mathcal{L}_P(r + G(x)_F, y_F)\right)$$

Everybody Dance Now: Results[14]



[14] Chan et al, Everybody Dance Now, ICCV 2019

Everybody Dance Now: Results[15]

Multi-subject synchronized dancing

[15] Chan et al, Everybody Dance Now, ICCV 2019

Homework

Readings

- Check this demo video from Everybody Dance Now paper
- Open Questions about Generative Adversarial Networks, Distill.pub
- (Optional) Papers on respective slides

Question

- Throughout this lecture, we saw methods that use videos as input for generating videos. Can we generate a video from a single image?

References

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. "Generating Videos with Scene Dynamics". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, 613–621.

J. Walker et al. "The Pose Knows: Video Forecasting by Generating Pose Futures". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 3352–3361.

C. Chan et al. "Everybody Dance Now". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 5932–5941.