Virtual Humans – Winter 23/24

Lecture 12_1 – Human Synthesis in a Scene

Prof. Dr.-Ing. Gerard Pons-Moll

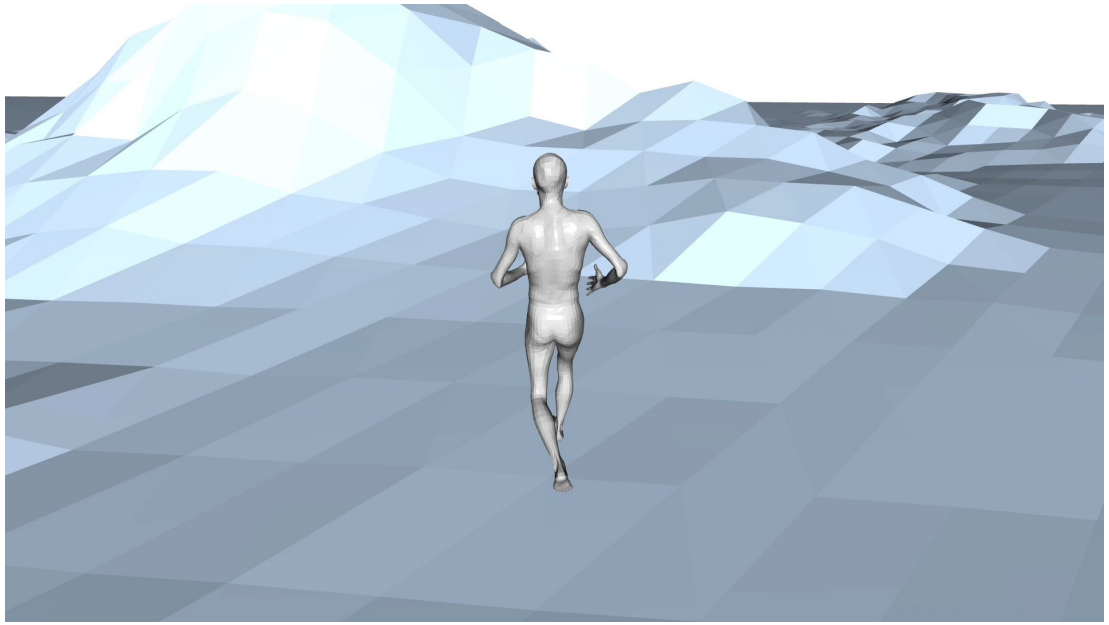
University of Tübingen / MPI-Informatics

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

In this lecture…

- Synthesising **static humans in static scenes**.
- Representing human-scene contacts.
- Refresher on generative models, VAE, cVAE.
- Synthesising **motion in an uneven terrain**.

Goal: **Awaken** Virtual Humans



Perceive: We should be able to reconstruct **real** 3D humans jointly with the objects and the scene they interact with

Generation: **Virtual** humans should be able to move and interact with objects and scenes like real humans

So far we have seen...

- We can capture human-object interaction (BEHAVE)
- We can reconstruct HOI from images (PHOSA, CHORE)

- **Can we also generate "human-object/scene" interaction?**
- **Why is "synthesis" of HOI useful?**


Gaming: GTA


Robotics: Agility
DIGIT
By Agility Robotoics


AR/VR: [Guzov et al 2021]


Metaverse: Meta
FACEBOOK
Briefing
Kurt
Welcom Scott
to
Worksrooms
Peter
Devindra
cnet

What do we need to synthesise HOI?

- We need to understand the 3D scene.
- Reason about **affordance**: A chair affords sitting, but also standing on it, grabbing it, etc
- Reason about **function**: The main function of a chair is to sit on it.
- We need to **synthesise 3D humans conditioned on the 3D scene**.

Can we synthesise **static 3D humans**, given a **static 3D scene**?

Hassan et al. CVPR'21

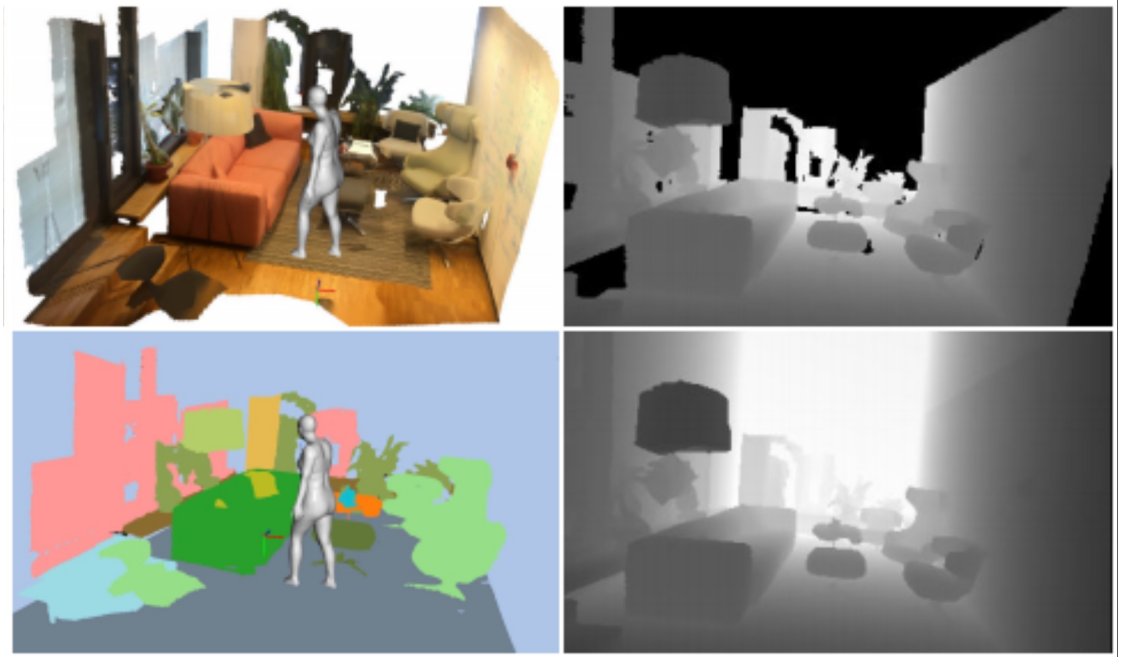
Is there data to learn such a model?

PROX (Hassan et al., ICCV 2021)

pseudo GT SMPL-X meshes in 3D scenes

PROX-E (Zhang et al., CVPR 2021)

semantic labels on top of PROX



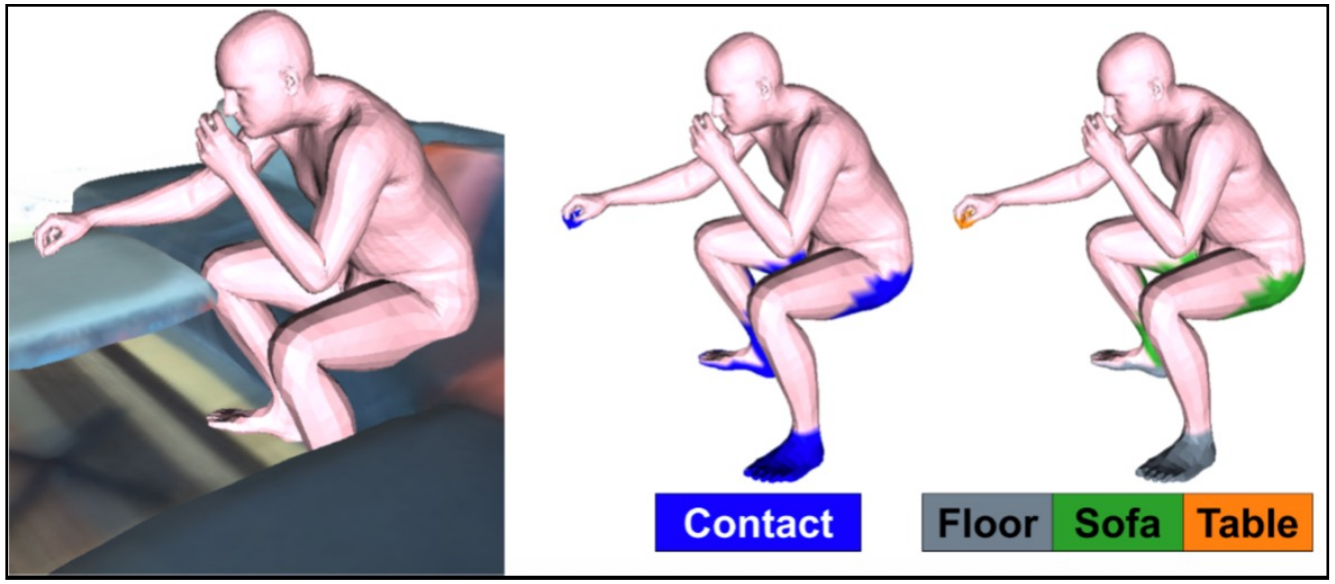
Problem: Current body models such as SMPL do not factor in the scene

To condition SMPL on scene, we need **contacts**

Key idea:

Based on the pose:

- Predict contact vertices.
- Predict likely objects in contact



Hassan et al. CVPR'21

Representing contacts, C

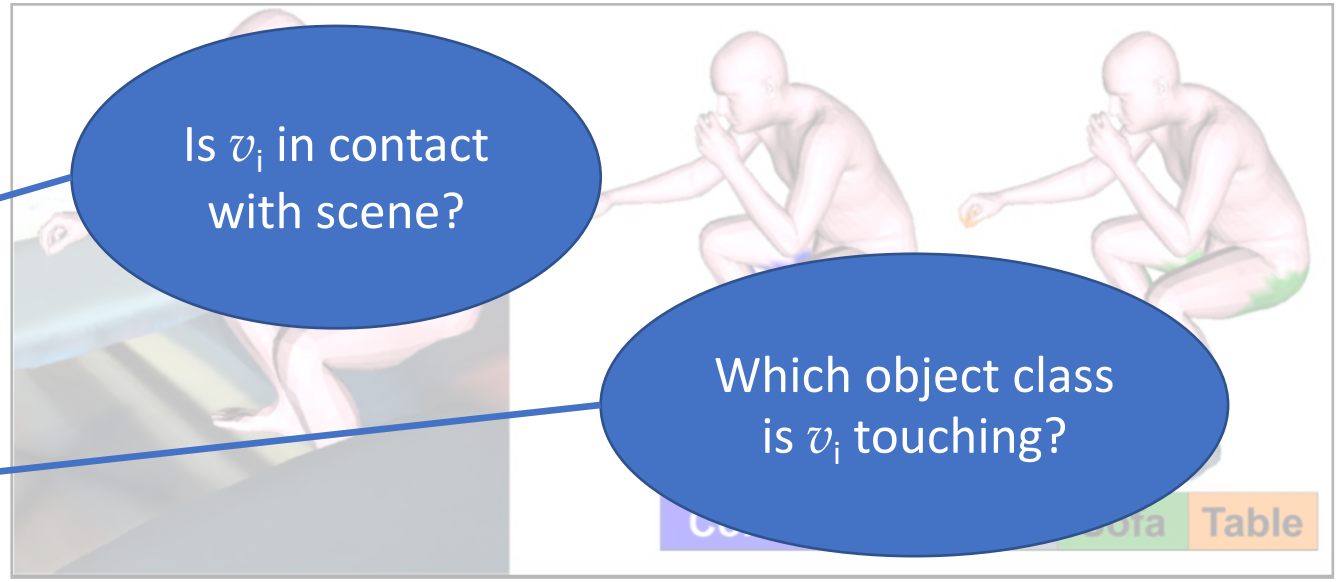
Body, scene vertices: V_b, V_s

$$C = \{[f_s, f_c]_i \ | v_i \in V_b\}$$

$$f_c \in \{0,1\}^{|V_b|}$$

$$f_s \in \{0,1\}^{|V_b| \times L}$$

L is the number of objects in scene.



Hassan et al. CVPR'21

Contacts can be sampled on SMPL vertices

Generative
Neural Network
e.g. VAE, cVAE

f_c

f_s

Hassan et al. CVPR'21

Slide credit: Siyu Tang

Refresher on VAE / cVAE

Variational Bayesian inference

The Bayes theorem

$$\underbrace{p(Z|X)}_{\text{Posterior}} = \frac{\overbrace{p(X|Z)}^{\text{Likelihood}} \cdot \overbrace{p(Z)}^{\text{Prior}}}{\underbrace{p(X)}_{\text{Marginal likelihood}}}$$

- X is the **evidence**, our observations from the system, the data.
- Z is the **hypothesis**, our assumptions on what causes the observations, the latent variables.
- The **likelihood** can represent the model of the system.
- In Bayesian inference, we calculate the **posterior** to infer the reasons.

Slide credit: Siyu Tang

Refresher on VAE / cVAE

Variational Bayesian inference

The Bayes theorem

$$\underbrace{p(Z|X)}_{\text{Posterior}} = \frac{\overbrace{p(X|Z)}^{\text{Likelihood}} \cdot \overbrace{p(Z)}^{\text{Prior}}}{\underbrace{p(X)}_{\text{Marginal likelihood}}}$$

- In most cases, the posterior does not have a closed form and is computationally intractable.
- Variational Bayesian inference uses another simpler distribution to approximate the posterior.
- **Two key questions:**
  - How to define the approximate posterior?
  - How to perform approximation?

Slide credit: Siyu Tang

Refresher on VAE / cVAE

Variational Bayesian inference

$$D_{KL}(q_\phi(Z|X)||p(Z|X)) = \int_Z q_\phi(Z|X) \cdot \log\left(\frac{q_\phi(Z|X)}{p(Z|X)}\right) dZ$$

Approximate posterior

Kullback-Leibler divergence

- The approximate posterior has a known form with unknown parameters.
- The Kullback-Leibler (KL) divergence measures the difference between two distributions.
  - Non-negative and convex
  - Non-symmetric measure
- To perform inference, we minimise the KL-divergence.

Slide credit: Siyu Tang

Refresher on VAE / cVAE

Variational autoencoding:

learn the generative model and approximate posterior simultaneously.

$$D_{KL}\left(q_\phi(Z|X)||p_\theta(Z|X)\right) \tag{1}$$

$$= \int_Z q_\phi(Z|X) \cdot \log\left(\frac{q_\phi(Z|X)}{p_\theta(Z|X)}\right) dZ \tag{2}$$

$$= \int_Z q_\phi(Z|X) \cdot \log\left(\frac{q_\phi(Z|X)p_\theta(X)}{p_\theta(X|Z)p_\theta(Z)}\right) dZ \tag{3}$$

$$= \log p_\theta(X) + \int_Z q_\phi(Z|X) \cdot \log\left(\frac{q_\phi(Z|X)}{p_\theta(Z)}\right) dZ - \int_Z q_\phi(Z|X) \cdot \log p_\theta(X|Z) dZ \tag{4}$$

$$= \log p_\theta(X) + D_{KL}\left(q_\phi(Z|X)||p_\theta(Z)\right) - E_{Z\sim q_\phi(Z|X)}[\log p_\theta(X|Z)] \tag{5}$$

Slide credit: Siyu Tang

Refresher on VAE / cVAE

Variational autoencoding:

learn the generative model and approximate posterior simultaneously.

$$\log p_\theta(X) - D_{KL}\left(q_\phi(Z|X)||p_\theta(Z|X)\right) = E_{Z\sim q_\phi(Z|X)}[\log p_\theta(X|Z)] - D_{KL}\left(q_\phi(Z|X)||p_\theta(Z)\right)$$

Maximise it for machine learning

Minimise it for variational inference

$$\log p_\theta(X) \geq E_{Z\sim q_\phi(Z|X)}[\log p_\theta(X|Z)] - D_{KL}\left(q_\phi(Z|X)||p_\theta(Z)\right)$$

Evidenced lower-bound (ELBO)

Slide credit: Siyu Tang

Refresher on VAE / cVAE

Variational autoencoding: evidenced lower-bound (ELBO) loss

The reconstruction term $E_{Z\sim q_\phi(Z|X)}[\log p_\theta(X|Z)]$

Encoding/inference **decoding/generation**

- Encoding is inference. Given a sample X, we derive the inference posterior and draw a latent variable Z.
- Decoding is generation. Given a latent variable Z, we generate a sample X'.
- Maximizing this term is equivalent to minimising the difference between X and X'.
- This term is used for data reconstruction. In practice, we can use L1 or L2 distance.

Slide credit: Siyu Tang

Refresher on VAE / cVAE

Variational autoencoding: evidenced lower-bound (ELBO) loss

The KLD term $D_{KL}\left(q_\phi(Z|X)||p_\theta(Z)\right)$

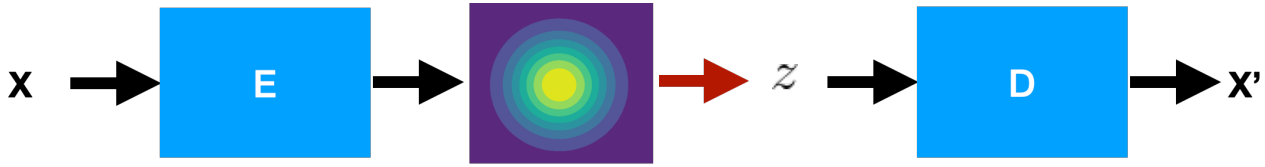
Inference posterior **Latent prior**

- The inference posterior has a known form with unknown parameters.
- The latent prior can be either pre-defined or learned from data.
- When both of them are Gaussian, the KLD term has a closed form.
- When this term is 0, the inference posterior is independent of X, leading to posterior collapse.

Refresher on VAE / cVAE

Variational autoencoding: the reparameterization trick

- We maximize the ELBO to train the VAE, via back-propagation.
- However, the sampling operation is non-differentiable.
- With re-parameterization, the gradients back-propagates without passing the sampling operation.



$$\frac{\partial x'}{\partial \phi} = \frac{\partial x'}{\partial z}\boxed{\frac{\partial z}{\partial \phi}}$$

?

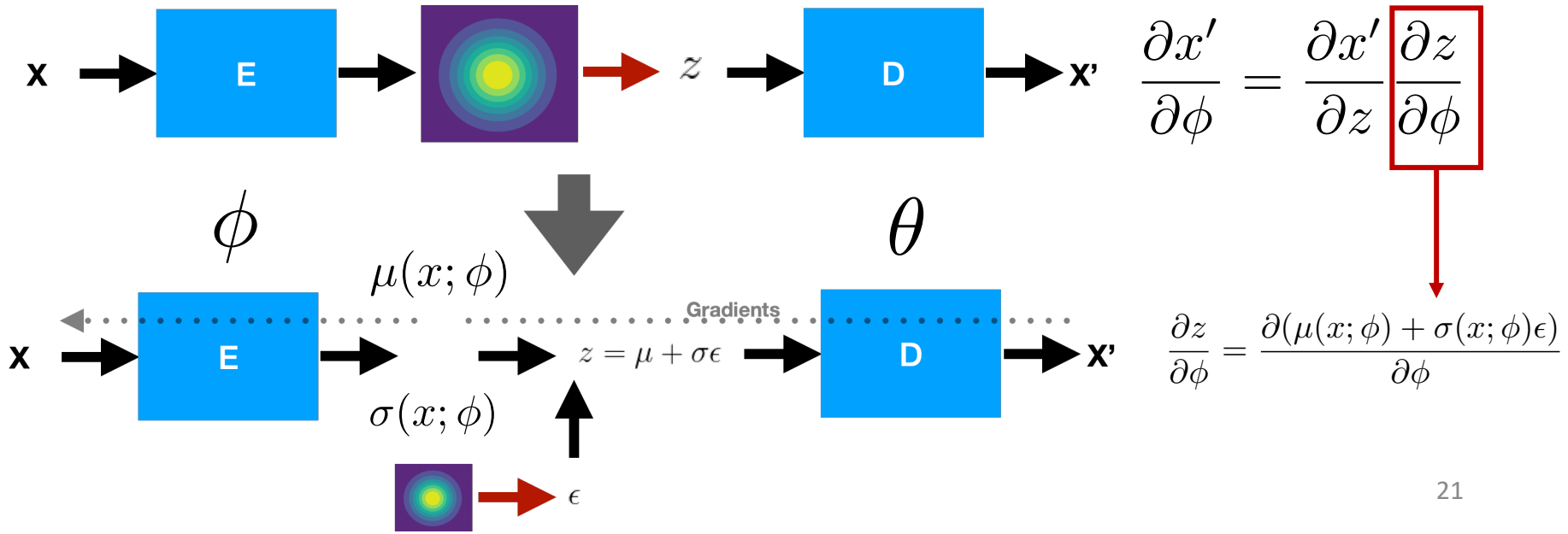
What I like the most about this paper!

Refresher on VAE / cVAE

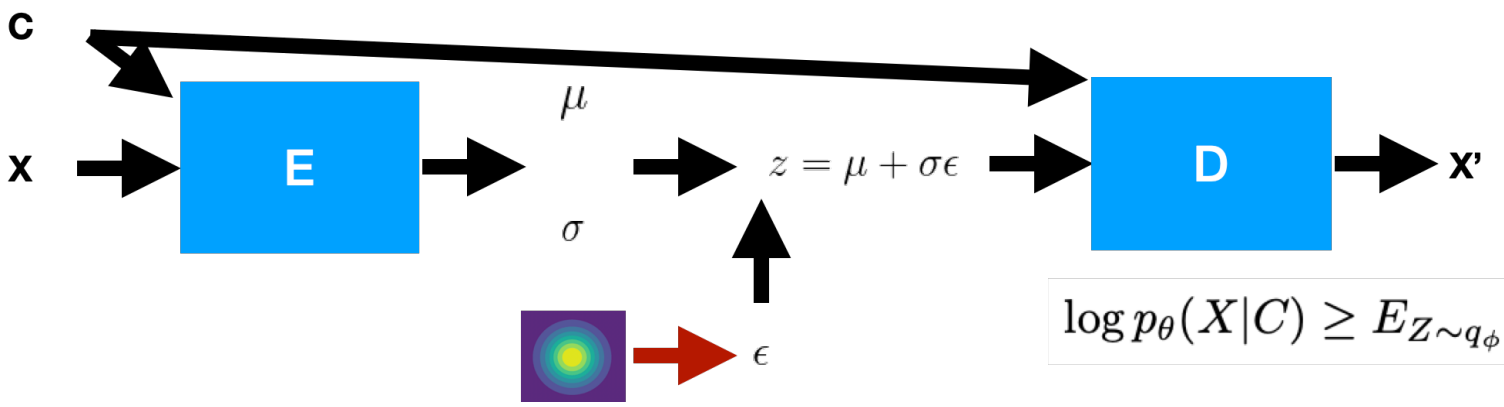
Variational autoencoding: the reparameterization trick

- We maximize the ELBO to train the VAE, via back-propagation.
- However, the sampling operation is non-differentiable.
- With re-parameterization, the gradients back-propagates without passing the sampling operation.

What I like the most about this paper!



Conditional VAE



$$\log p_\theta(X|C) \geq E_{Z\sim q_\phi(Z|X,C)}[\log p_\theta(X|Z,C)] - D_{KL}\left(q_\phi(Z|X,C)||p_\theta(Z)\right)$$

- The condition '*c*' can be action labels, motions from the past, or the scene context.
- The condition is concatenated with both the encoder and the decoder.
- The cVAE is widely used for motion modelling.

Contacts can be sampled on SMPL vertices

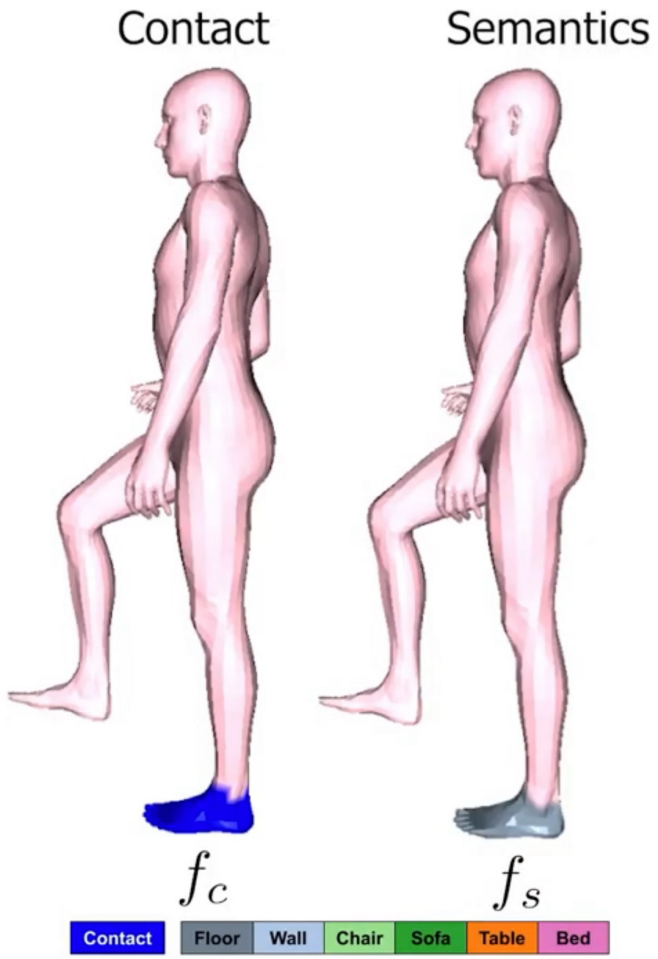
Generative
Neural Network
e.g. VAE, cVAE

f_c

f_s

Hassan et al. CVPR'21

Sampled contacts using cVAE



Hassan et al. CVPR'21

Fitting SMPL to scene using sampled contacts

Optimise SMPL pose and translation

$$E(\theta, t) = L^{\text{afford}} + L^{\text{pen}} + L^{\text{reg}}$$

$$L^{\text{afford}} = ||f_c \cdot f_d||_2 + \lambda \sum_i CCE(f_s^i, f_{ds}^i)$$

f_c Target contacts predicted by NN

f_d Distance between current SMPL and scene

f_s Target object predicted by NN

f_{ds} Current contacting object

Fitting SMPL to scene using sampled contacts

Optimise SMPL pose and translation

$$E(\theta, t) = L^{\text{afford}} + L^{\text{pen}} + L^{\text{reg}}$$

$$L^{\text{pen}} = \sum_{f_d^i < 0} (f_d)^2$$

f_d Signed distance between current SMPL and scene

Fitting SMPL to scene using sampled contacts

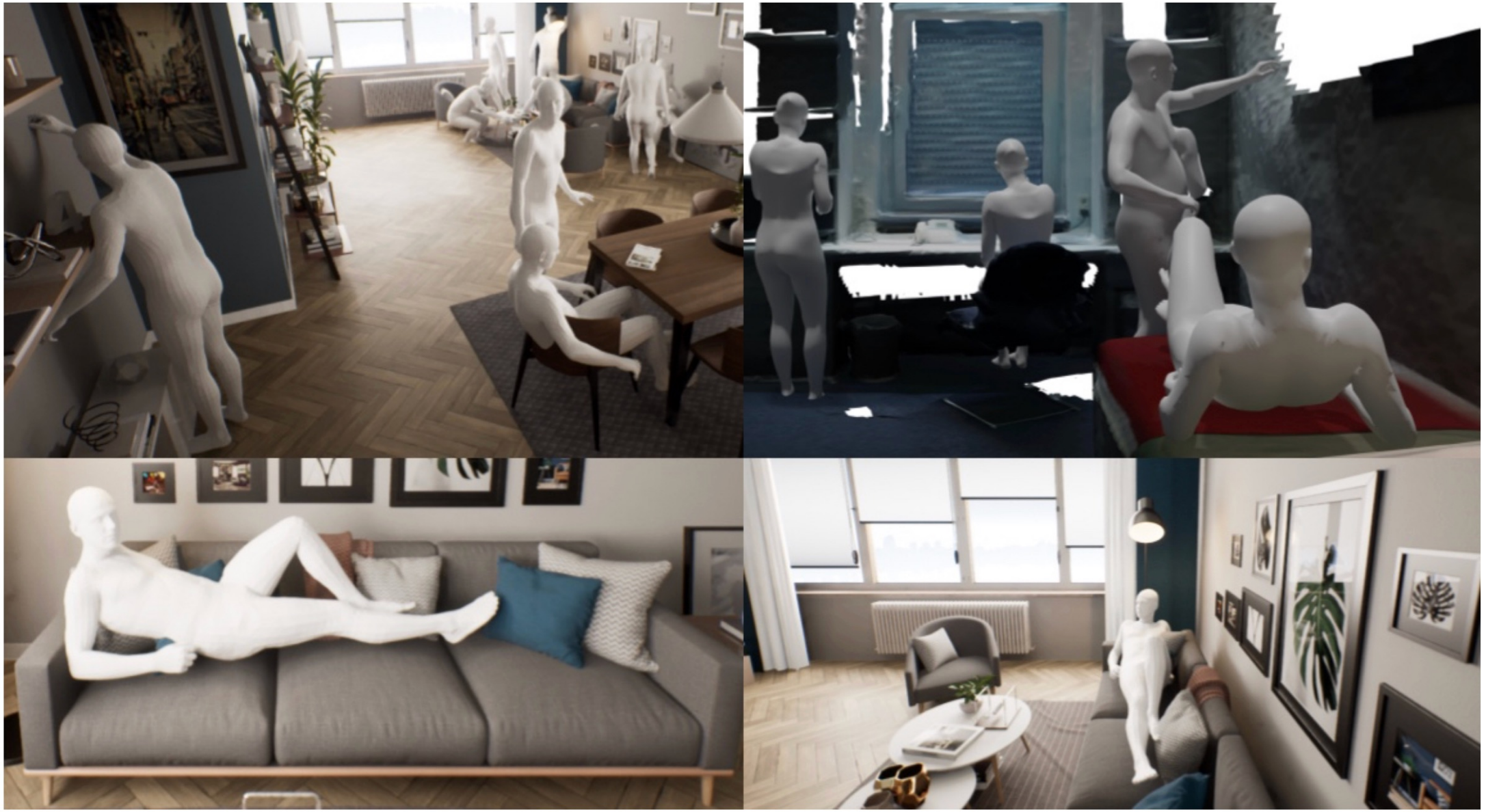
Optimise SMPL pose and translation

$$E(\theta, t) = L^{\text{afford}} + L^{\text{pen}} + L^{\text{reg}}$$

$$L^{\text{reg}} = ||\theta - \theta_{\text{init}}||^2$$

Current pose should not deviate too much from initialisation.

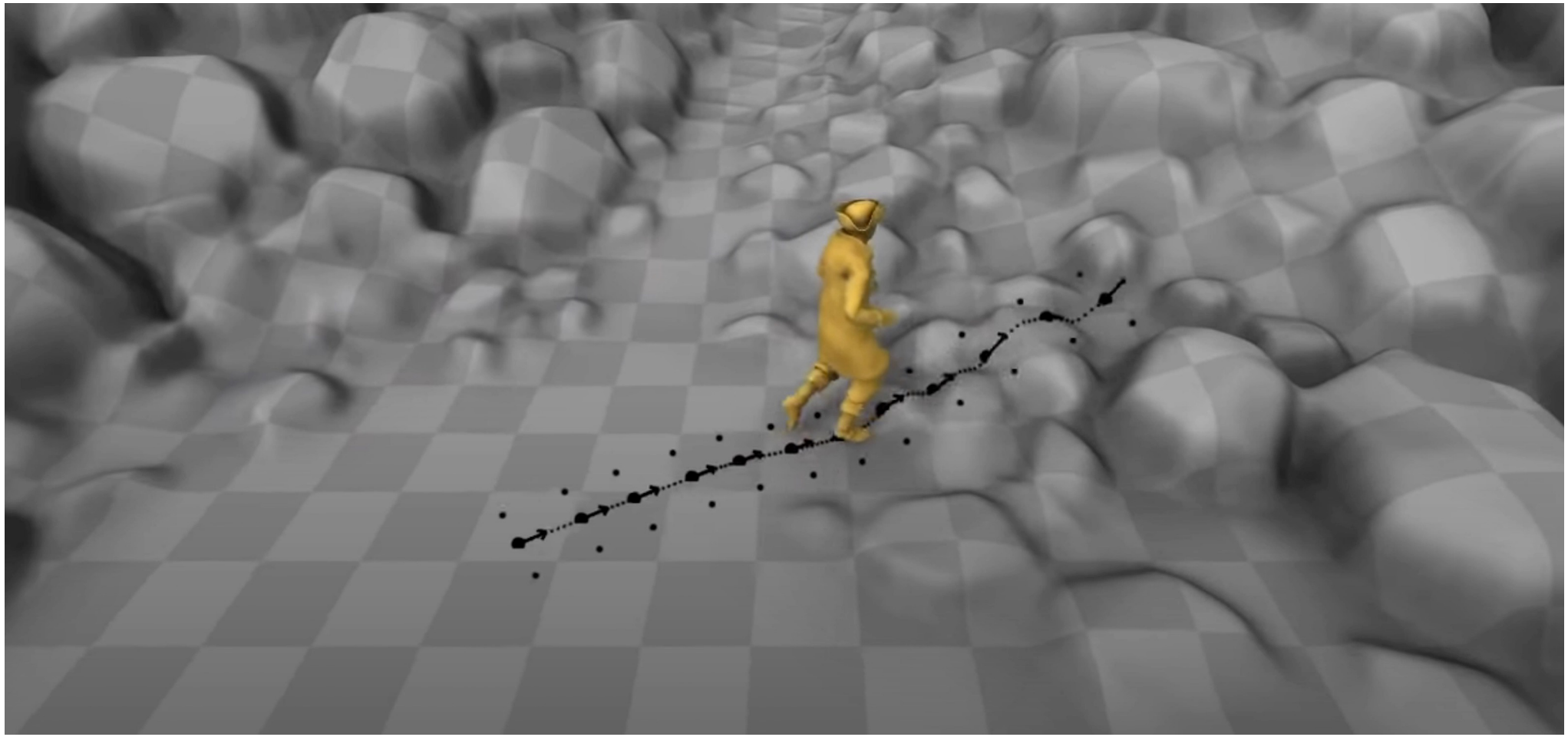
We can optimise **static SMPL** conditioned on scene



What about dynamic poses?

That is a much harder problem, let’s dive into it!

Given 3D terrain and type of motion, synthesize a sequence of 3D poses



PFNN. Holden et al. SIGGRAPH'17

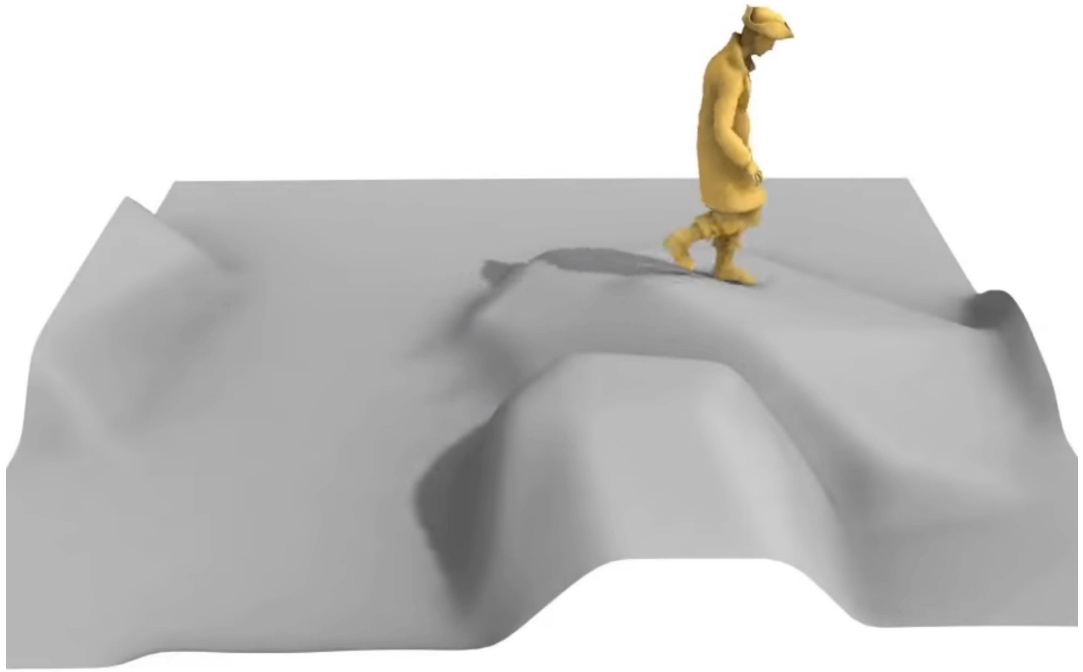
Challenges

1. How to obtain data to learn such a model?
2. How do we encode the motion and terrain?
3. How do we perform inference?

PFNN. Holden et al. SIGGRAPH'17

1. How to obtain data to learn such a model?

- Capturing data with varying terrain is hard.
- Record a subject walking and climbing stairs/ stool.
- Optimize the the terrain to fit the captured motion.



PFNN. Holden et al. SIGGRAPH'17

2. How to encode motion and terrain?

$$J_t, J'_t, T_t, T'_t, H_t, \Delta\phi = f(J_{t-1}, J'_{t-1}, T_{t-1}, T'_{t-1}, H_{t-1}, \phi_t)$$

Joint locations

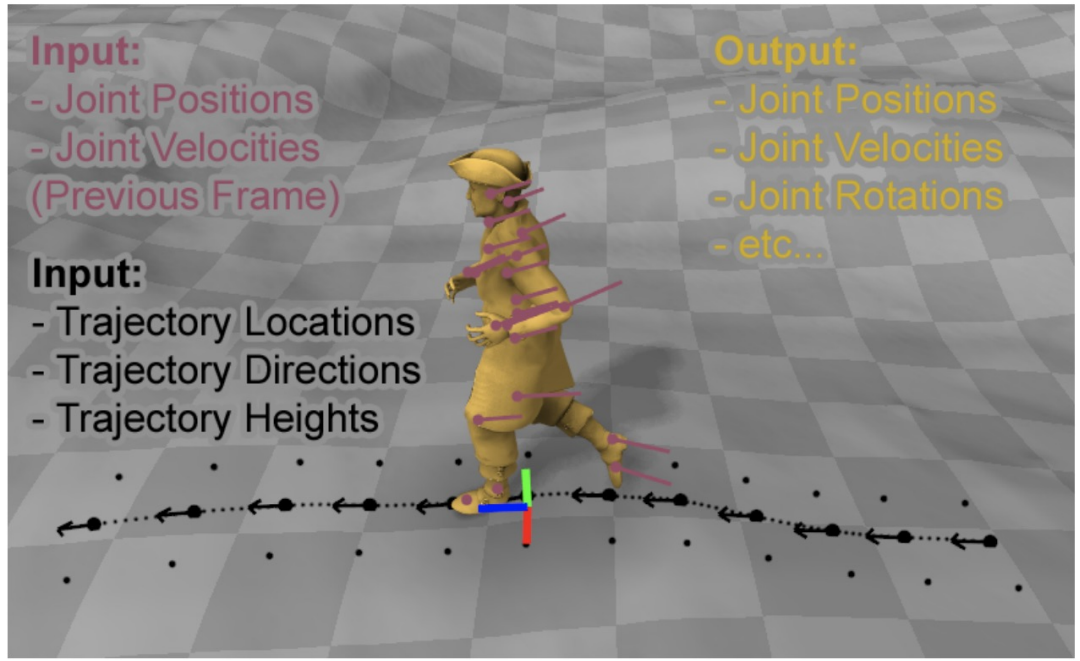
Joint velocity

Root Trajectory (xy-location)

Root Velocity

Height

Phase

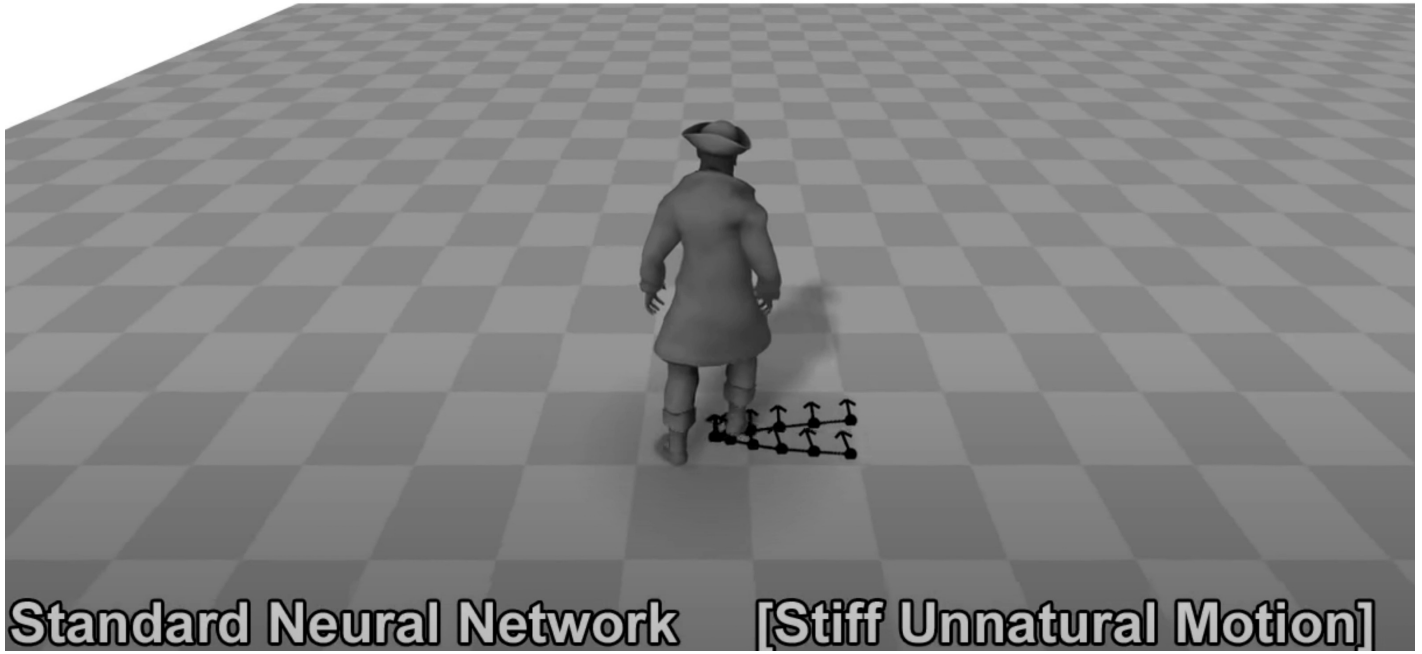


- The motion is encoded as location and velocity of root and joints.
- The terrain is encoded as the height at sampled points around trajectory.

PFNN. Holden et al. SIGGRAPH'17

What is phase, Φ?

- The phase Φ is an auxiliary variable that cycles between 0 and 2π.
- It represents the progress of one walking cycle,
  e.g. 0 could be a foot lift off and 2π is when we again land the foot.
- Without phase, motion is stiff and has foot sliding artefact.



PFNN. Holden et al. SIGGRAPH'17

3. How to predict future motion?

$$J_t, J'_t, T_t, T'_t, H_t, \Delta\phi = f(J_{t-1}, J'_{t-1}, T_{t-1}, T'_{t-1}, H_{t-1}, \phi_t)$$

Joint locations

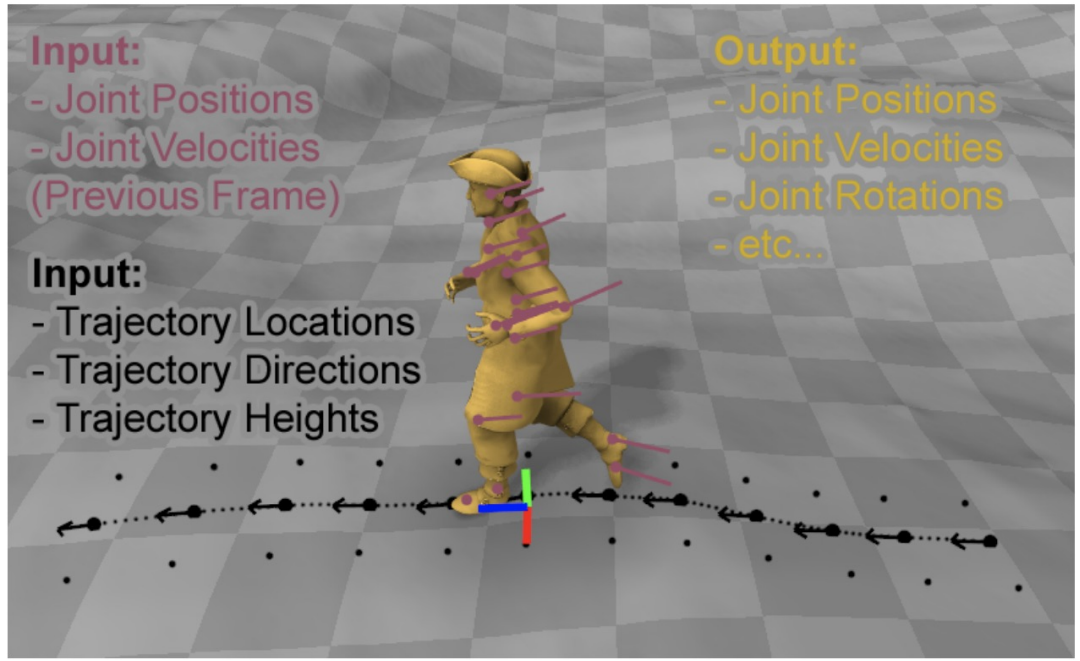
Joint velocity

Root Trajectory (xy-location)

Root Velocity

Height

Phase



- Future motion is predicted auto-regressively as a function of past predictions.
- f(.) is a neural network with MLPs.

PFNN. Holden et al. SIGGRAPH'17

Synthesised motion on challenging terrain

PFNN. Holden et al. SIGGRAPH'17

Takeaways

- Current models like SMPL do not model humans as a function of scene.
- To synthesise humans in static scene, we need contacts.
- Contacts encode which body point touches which scene point.
- One way to generate contacts is using generative modelling e.g. VAE.
- Motion synthesis can me modelled as as a auto-regression task with past motion and terrain as input.