Virtual Humans – Winter 23/24

Lecture 7_2 – Fitting SMPL to IMU with learning

Prof. Dr.-Ing. Gerard Pons-Moll
University of Tübingen / MPI-Informatics

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

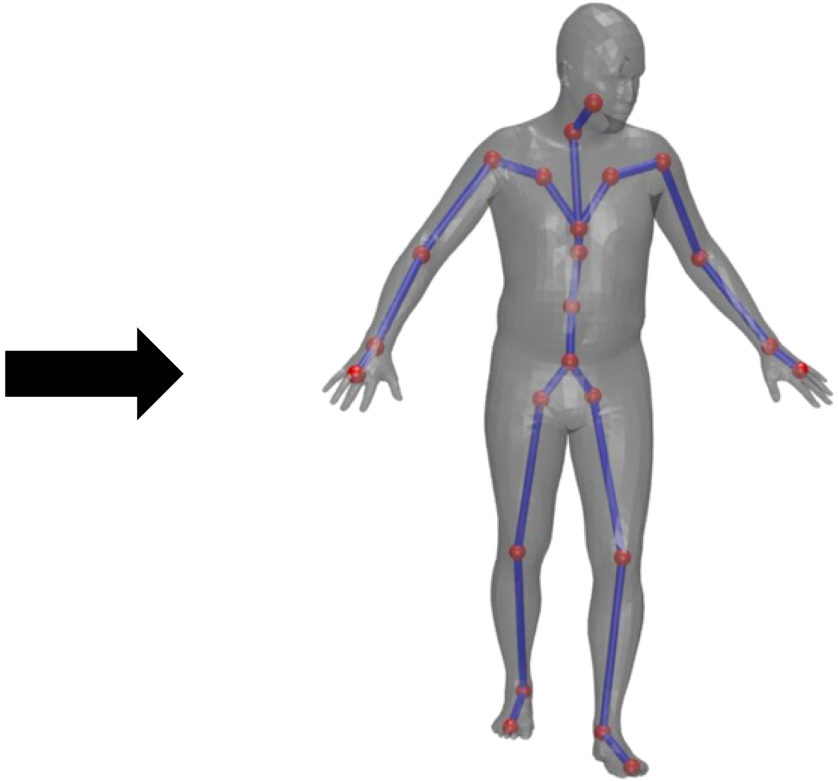
Deep Inertial Poser

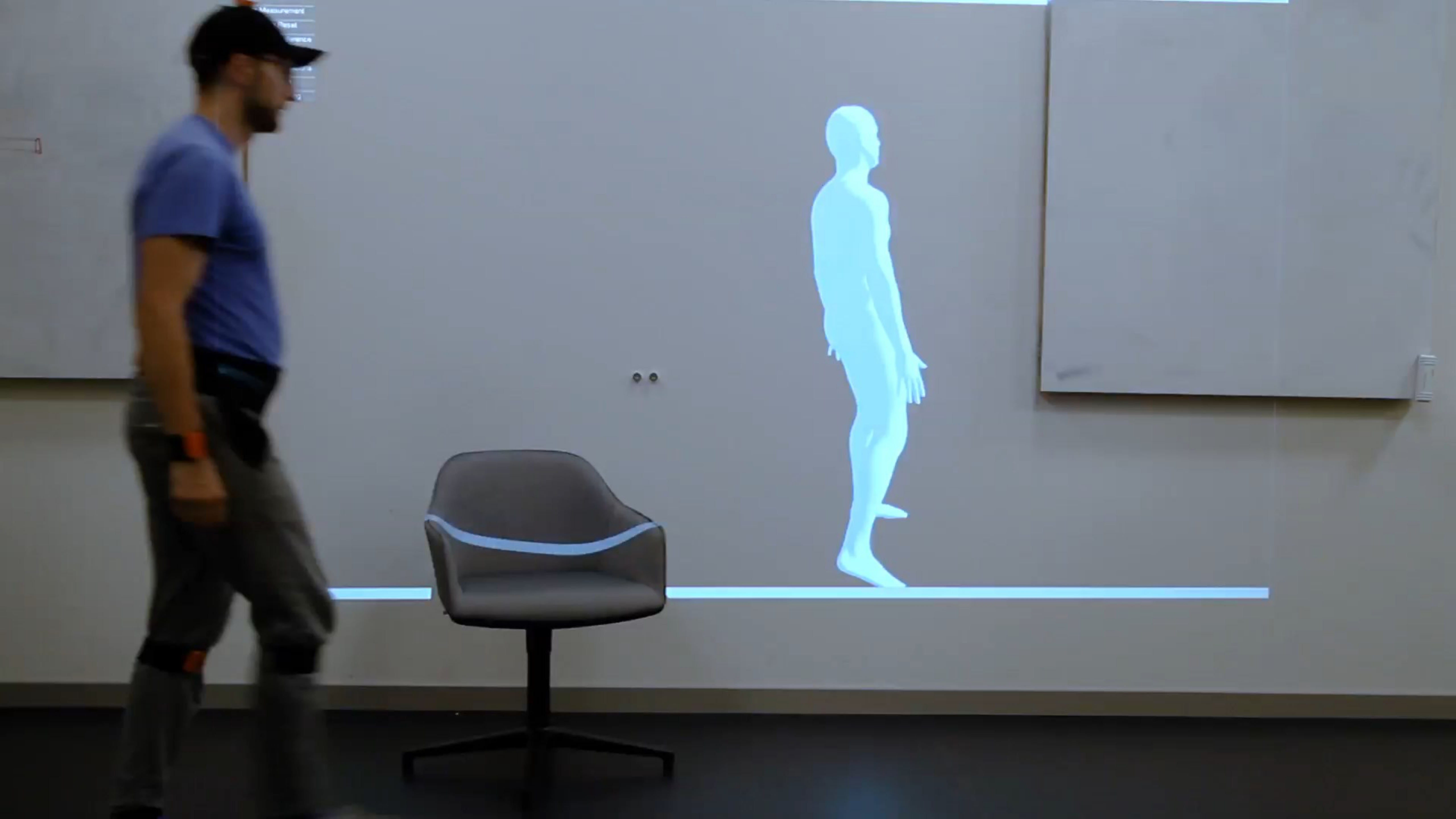
Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time

Yinghao Huang*[1], **Manuel Kaufmann***[2], Emre Aksan[2], Michael J. Black[1], Otmar Hilliges[2], Gerard Pons-Moll[3]

* equal contribution, [1] MPI for Intelligent Systems, [2] ETH Zurich, [3] MPI for Informatics

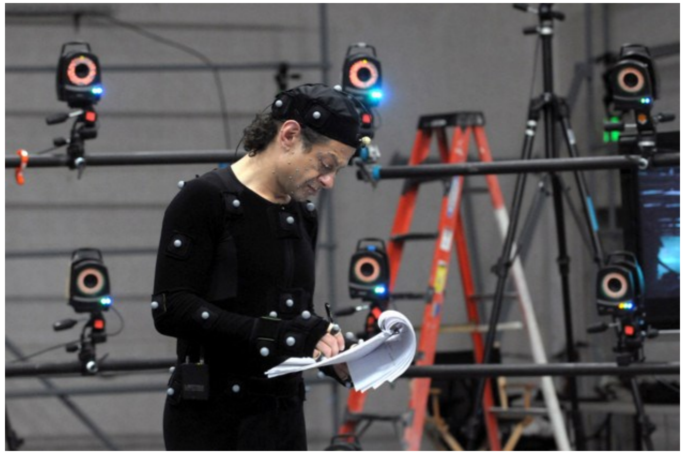
Goal




Measurement
Reset

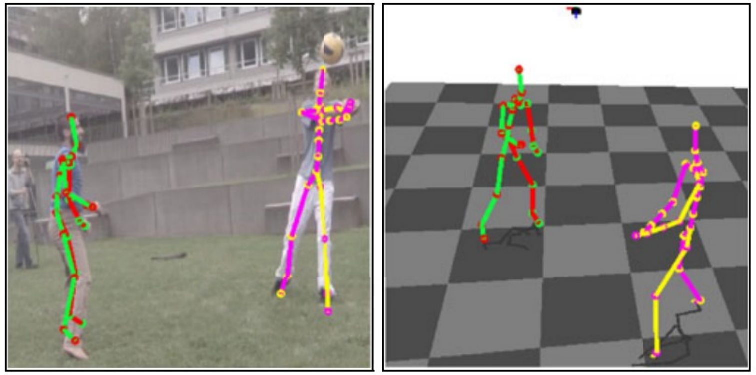
Motion Capture – Optical Tracking

Marker-based



- Long setup times
- Expensive equipment

Markerless



[Elhayek et al. 2017, MARCOnI]

[Mehta et al. 2017, VNect]

- Fixed recording volume
- Requiring line of sight

Motion Capture – Inertial Sensors

Number of IMUs

[Roetenberg et al. 2007]

- Intrusive
- 17 sensors

Cameras

[Malleson et al. 2017]

[von Marcard et al. 2018]

- 6 – 13 sensors
- 1 – 8 cameras

Compute Time

[von Marcard et al. 2017]

- offline

DIP - Requirements

Small number of IMUs

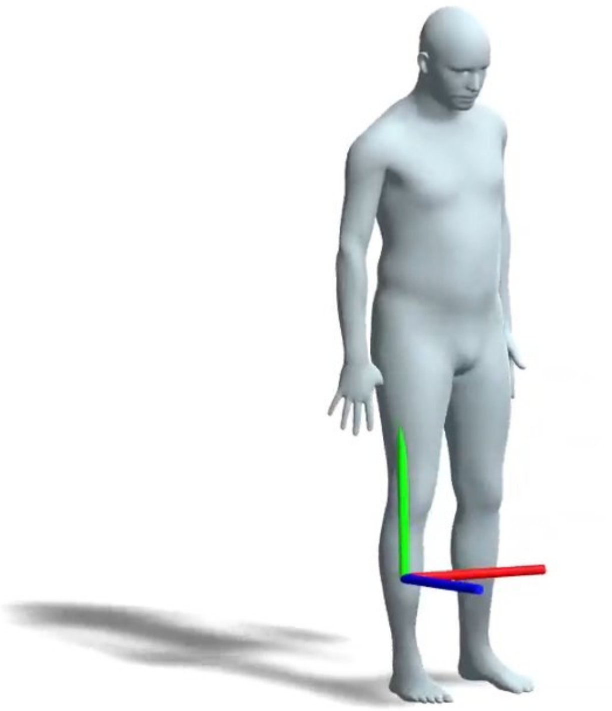
(setup time, user instrumentation)

No cameras

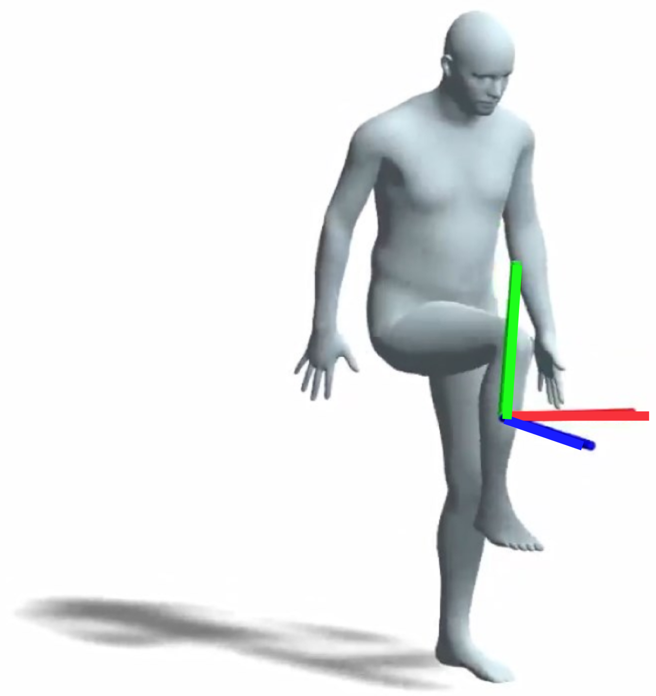
(line-of-sight, occlusions)

Reconstruct full pose in **real-time**

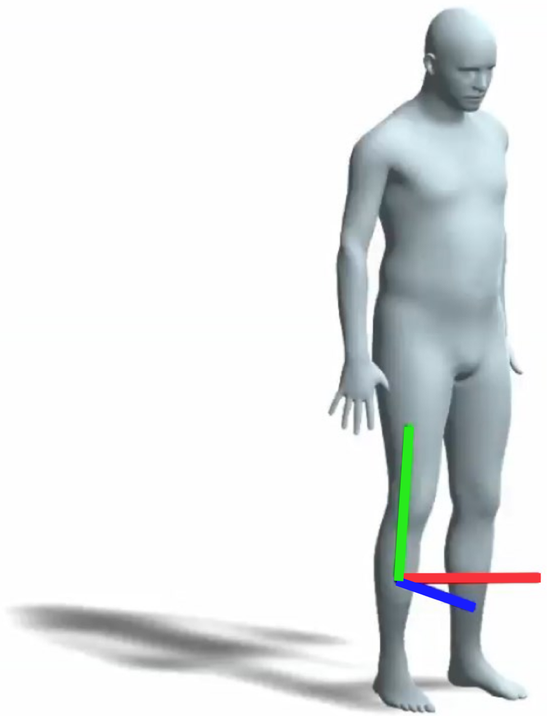
Underconstrained Pose Space



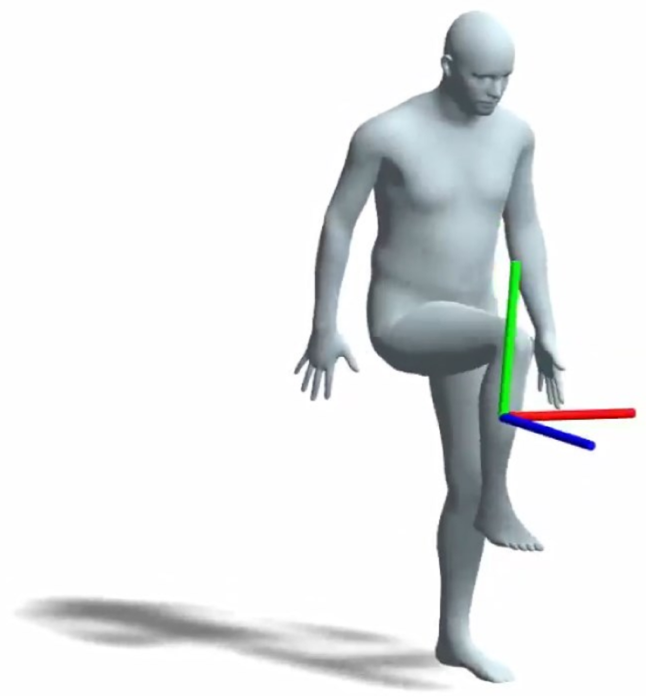
Underconstrained Pose Space



pose
different

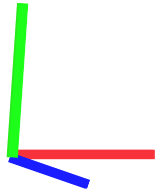


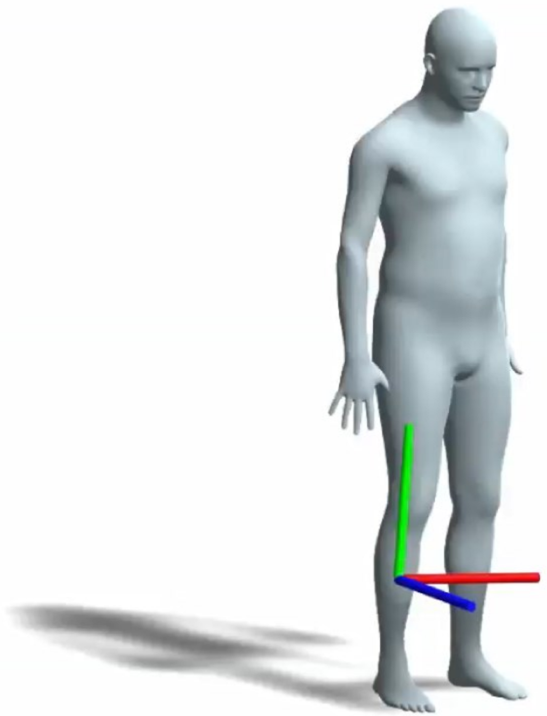
Underconstrained Pose Space



pose
different

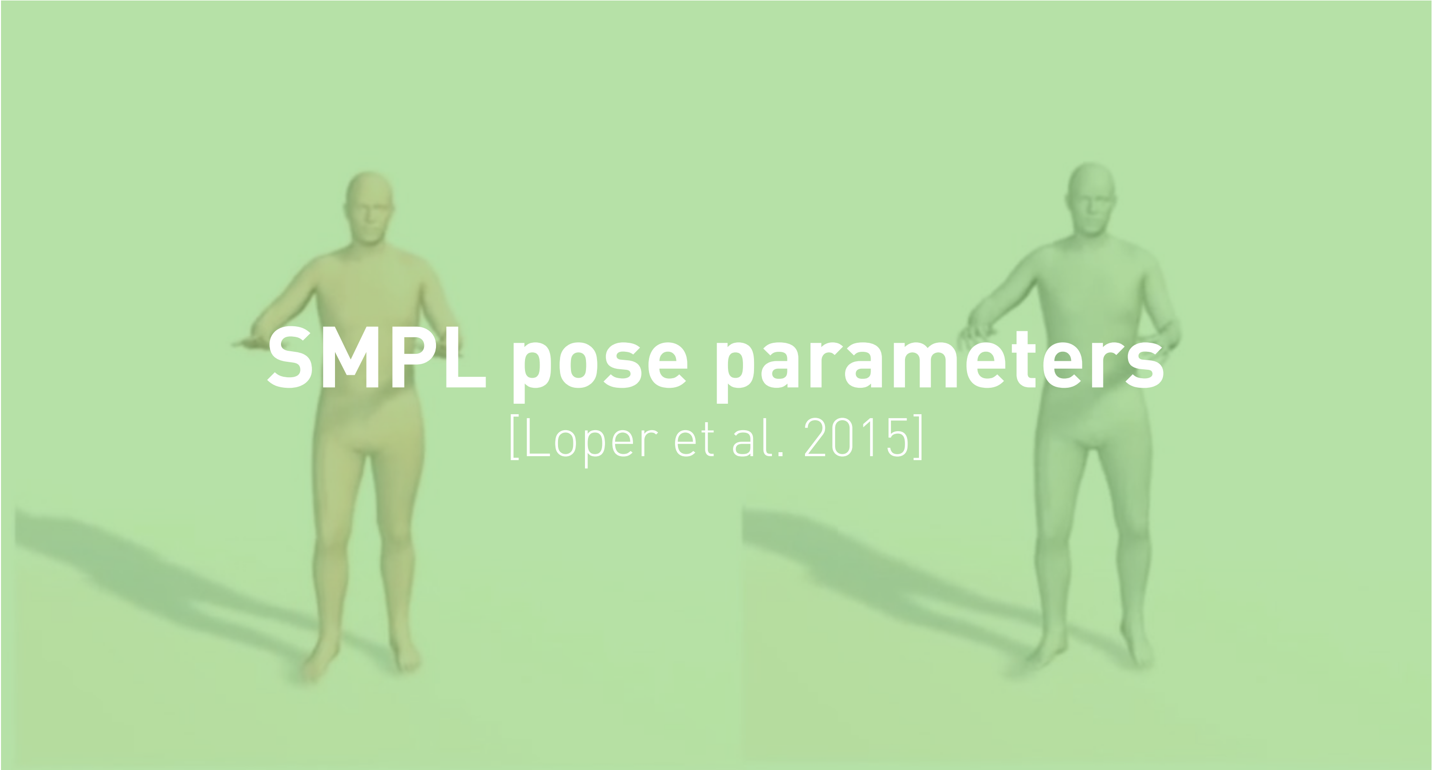
orientation
measurement
nearly identical





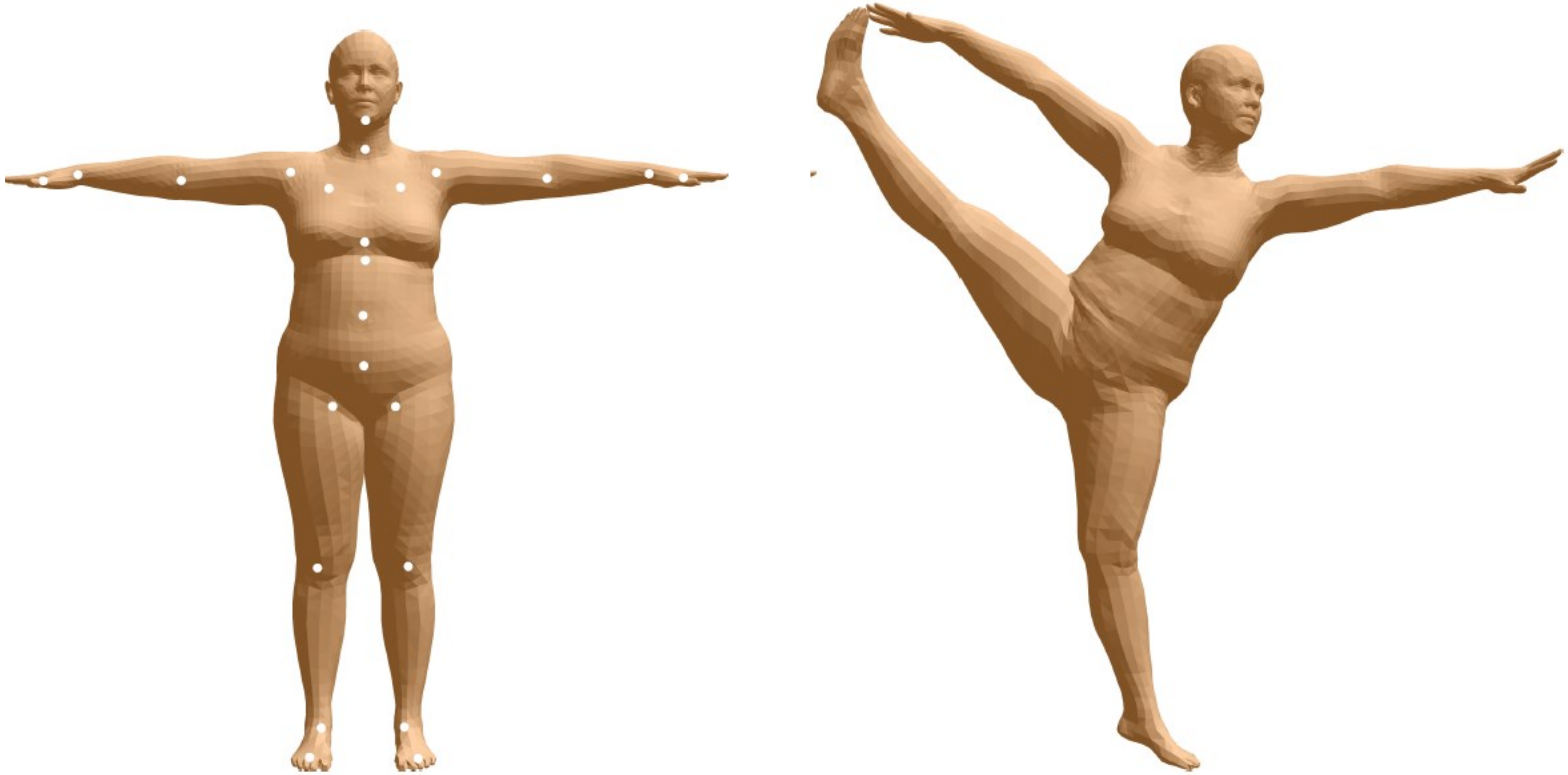
Sparse Inertial Poser (SIP)

[von Marcard et al. 2017]



SMPL

[Loper et al. 2015]

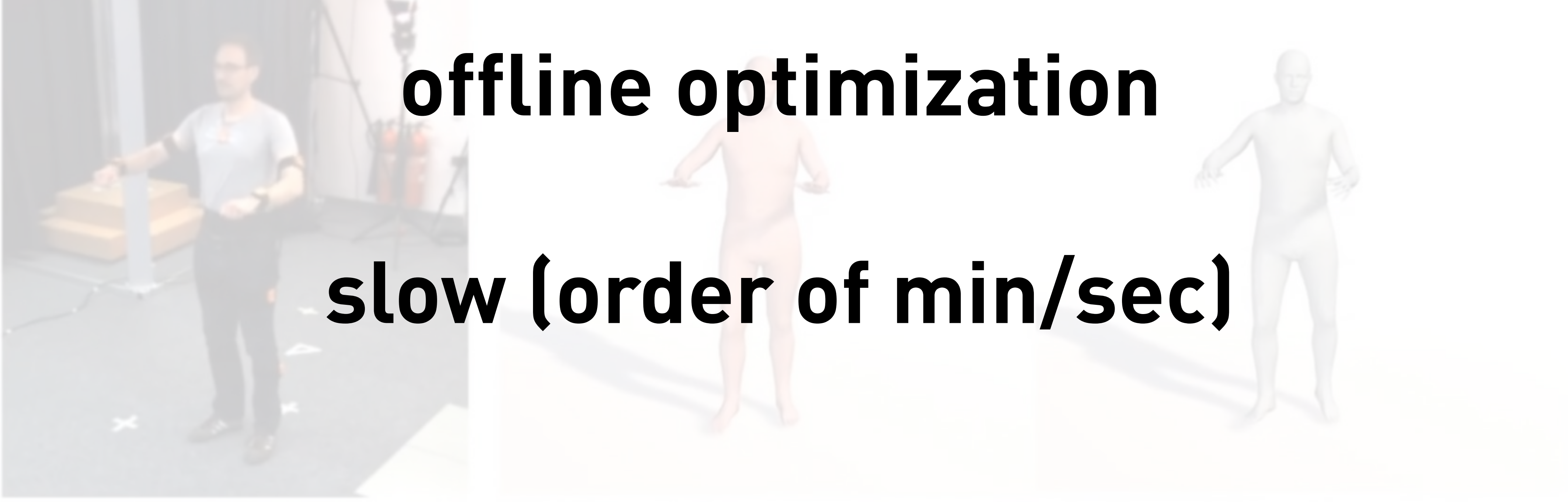


Sparse Inertial Poser (SIP)

[von Marcard et al. 2017]

SOP
orientation only

SIP
orientation + acceleration

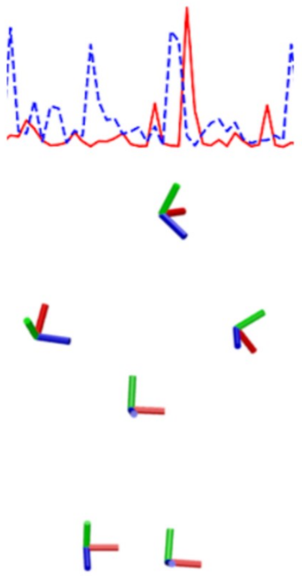


offline optimization

slow (order of min/sec)

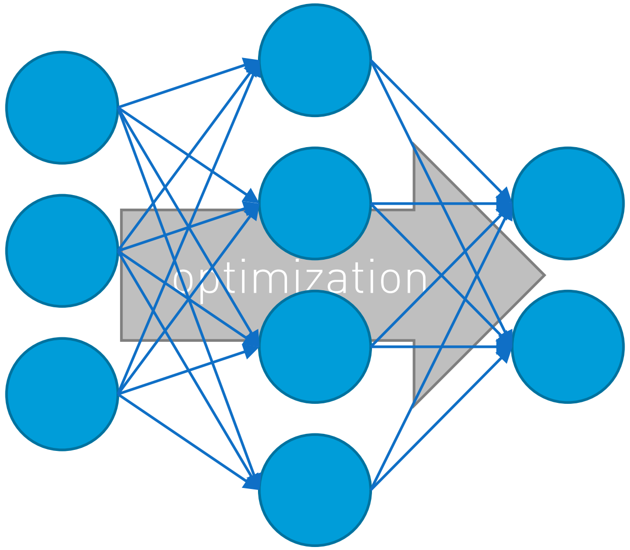
Achieving Real-Time Performance

Data

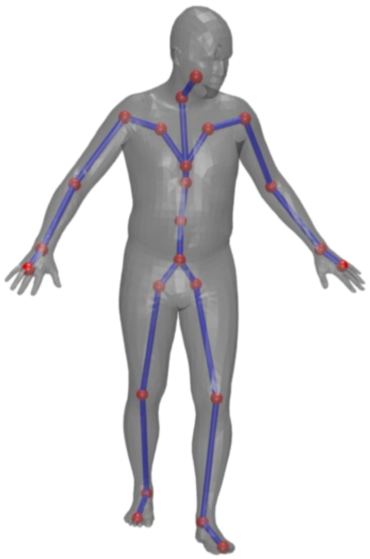


accelerations and orientations

Architecture



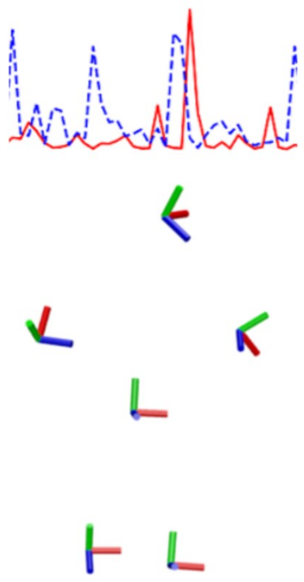
Loss Function



SMPL
pose parameters

Achieving Real-Time Performance

Data



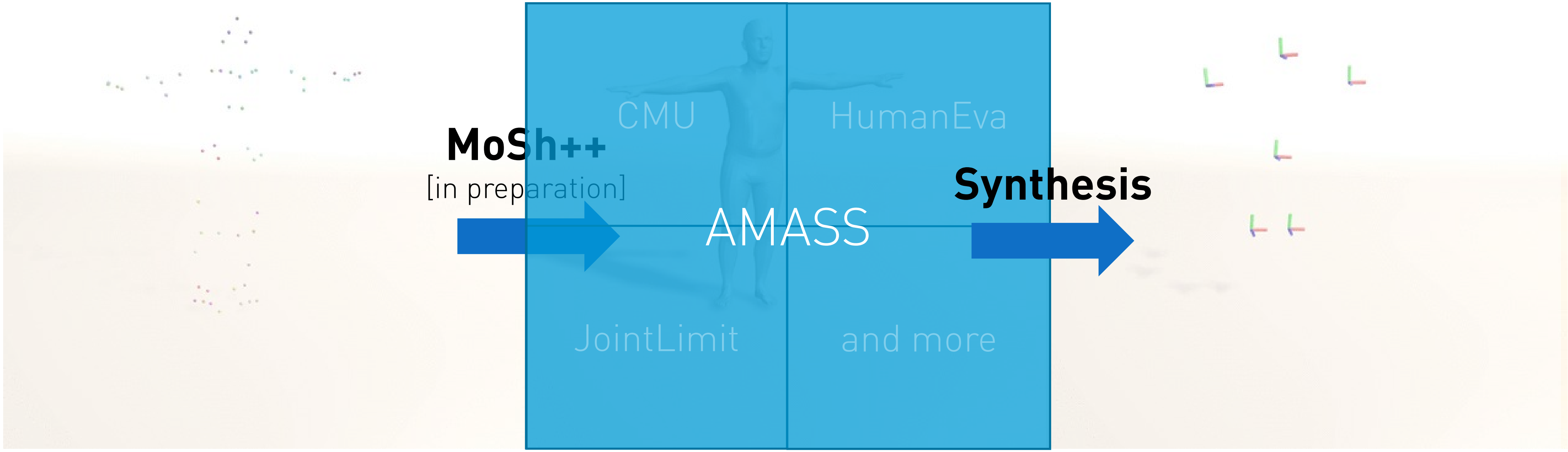
accelerations and orientations

How to Get Data?

Only **few** IMU databases available.

Need poses in **unified format**.

Synthesize It!



Acceleration
(derived from positions via finite differences)

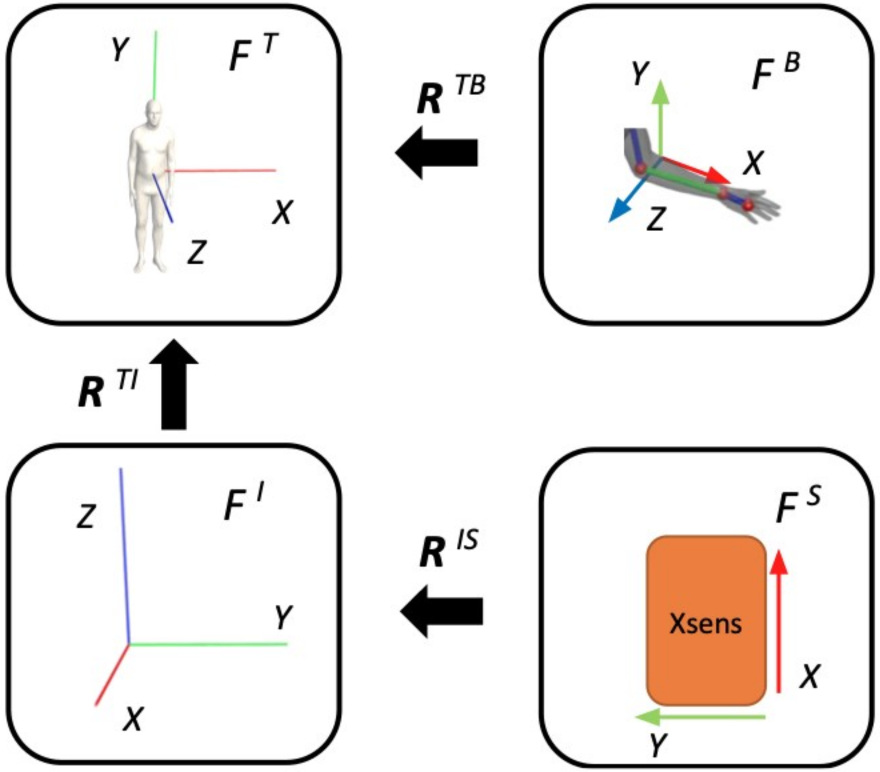
http://dip.is.tue.mpg.de

Orientation
(derived from SMPL forward kinematics)

Coordinate frames involved

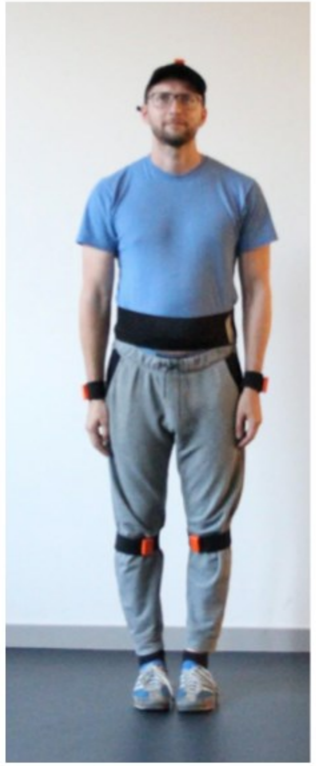
Body Centric frame

Bone frame
(body part)



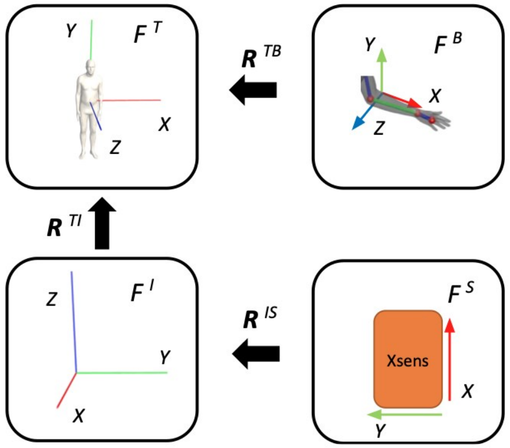
Frame of IMU
system (inertial)

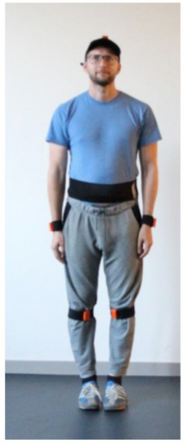
Local sensor frame

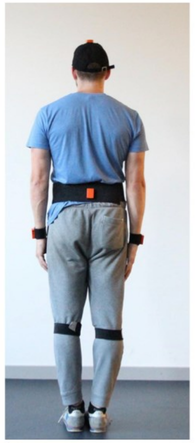


Orientation

1) IMU readings need to be transformed to body coordinate frame F^T







$$\mathbf{R}_t^{TS} = \mathbf{R}^{TI}\mathbf{R}_t^{IS} = \mathbf{R}_{\text{Head}}^{-1}\mathbf{R}_t^{IS}$$

Head sensor aligned with body in frame 0

2) Compensate for an assumed constant sensor to body part / bone offset

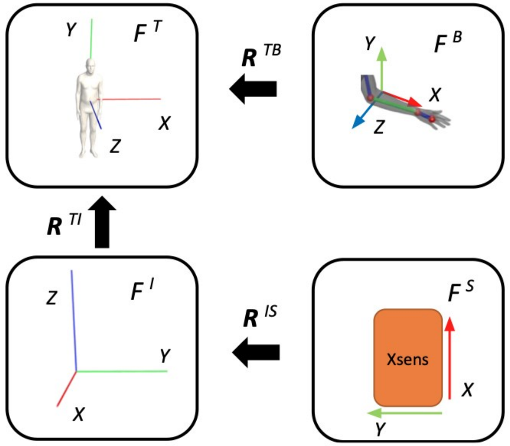
$$\mathbf{R}^{BS} = \text{inv}(\mathbf{R}_0^{TB})\mathbf{R}_0^{TS}$$

Sensor to bone offset calculation, usually in the frame 0

$$\mathbf{R}^{TB} = \mathbf{R}^{TS}\text{inv}(\mathbf{R}^{BS})$$

Transform IMU reading to bone orientations

Orientation



Question: what problem do you foresee if we train a network directly to predict pose from bone transformations as described below?

Hint: Think of a motion performed facing north vs facing south

Normalization

Normalize all sensors to the **root** sensor.

Done **per frame**.

Only **5 sensors** are actually fed into the model.

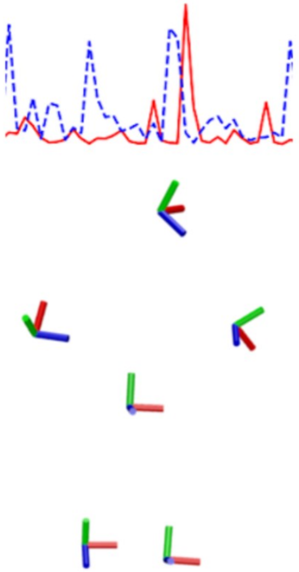
$$\mathbf{R}_t^{TB} = \mathbf{R}^{BS}\mathbf{R}_t^{TS},$$

normalize

$$\bar{\mathbf{R}}_t^{TB} = \mathrm{inv}(\mathbf{R}_{\mathrm{t}}^{\mathrm{root}})\mathbf{R}_t^{TB}$$

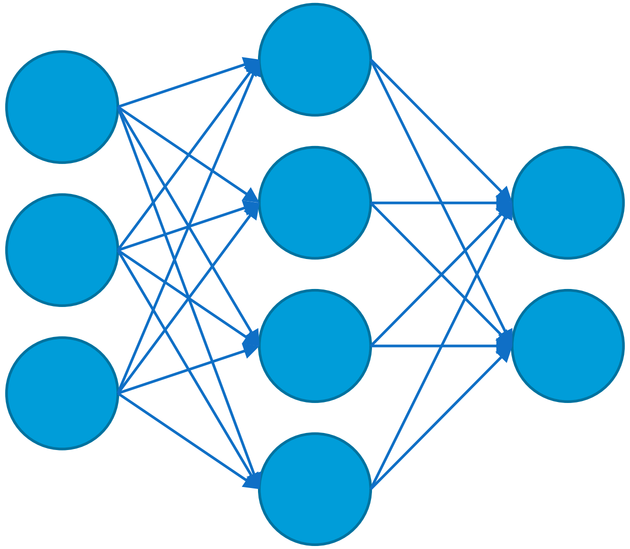
Network Design

Data

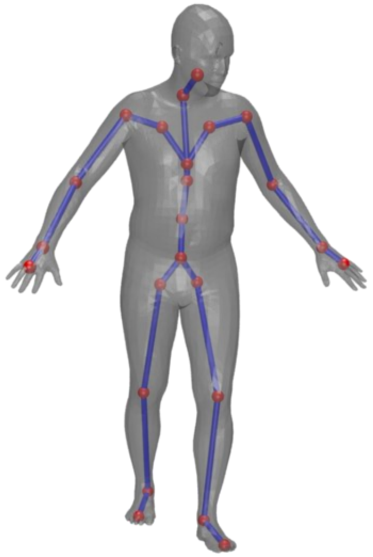


accelerations and orientations

Architecture

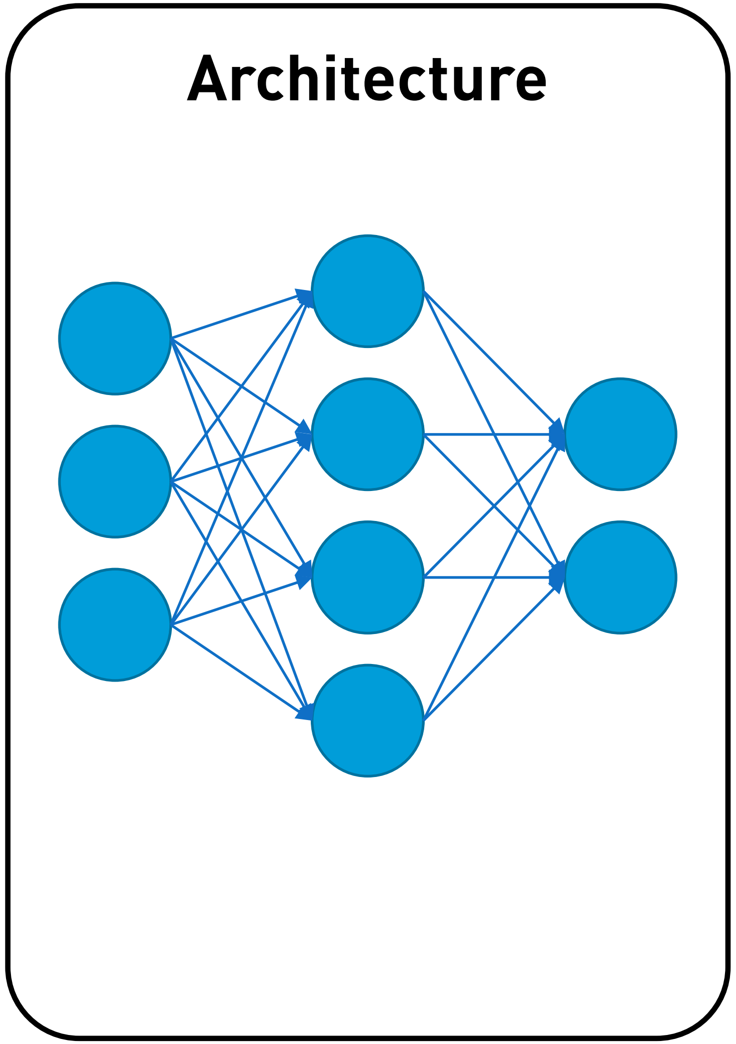


Loss Function

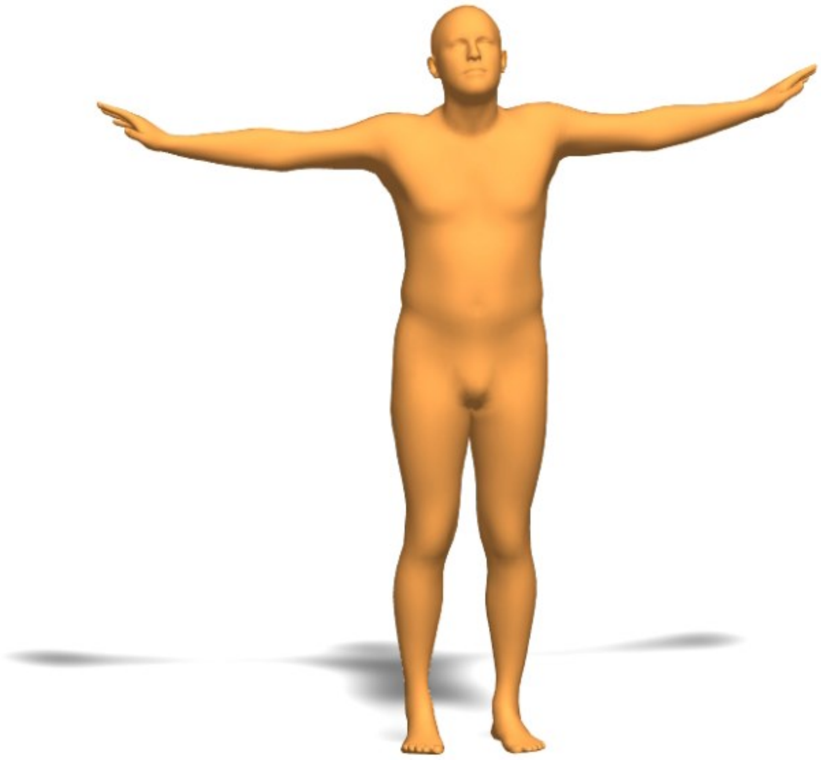


SMPL pose parameters

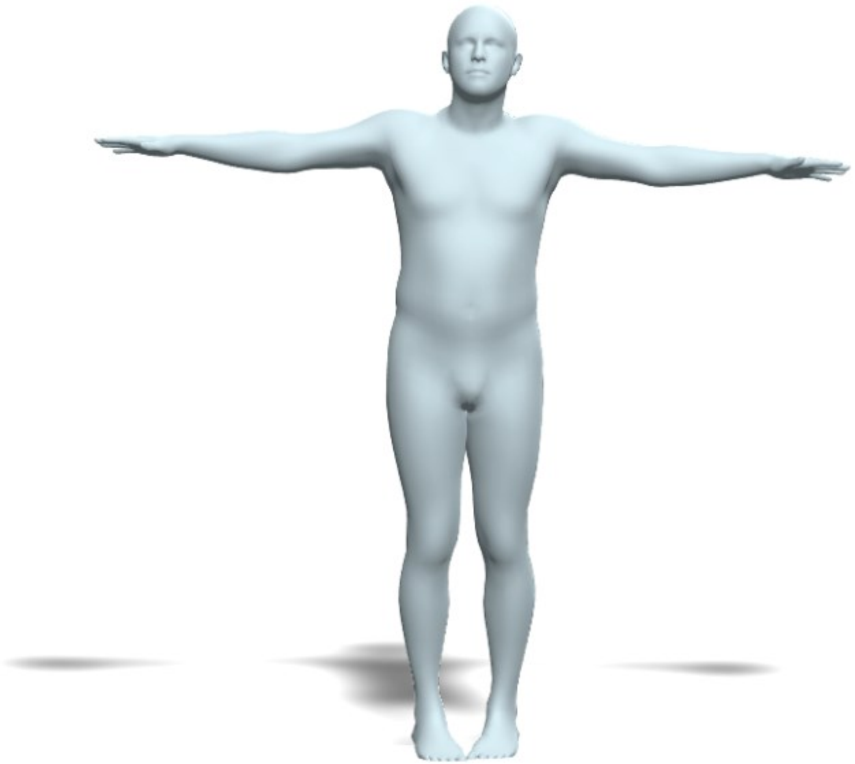
Network Design



Failed Attempt I

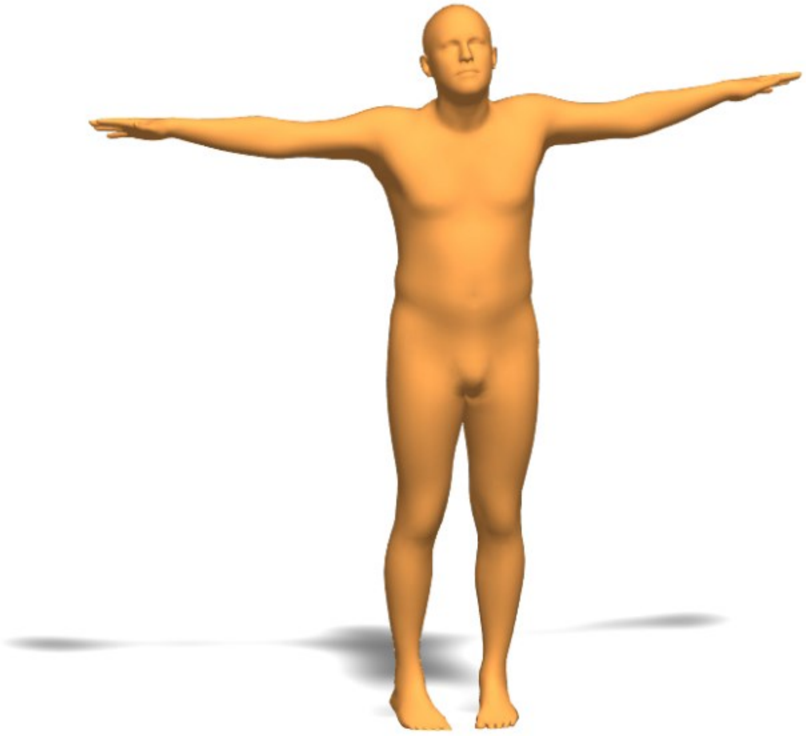


Reference

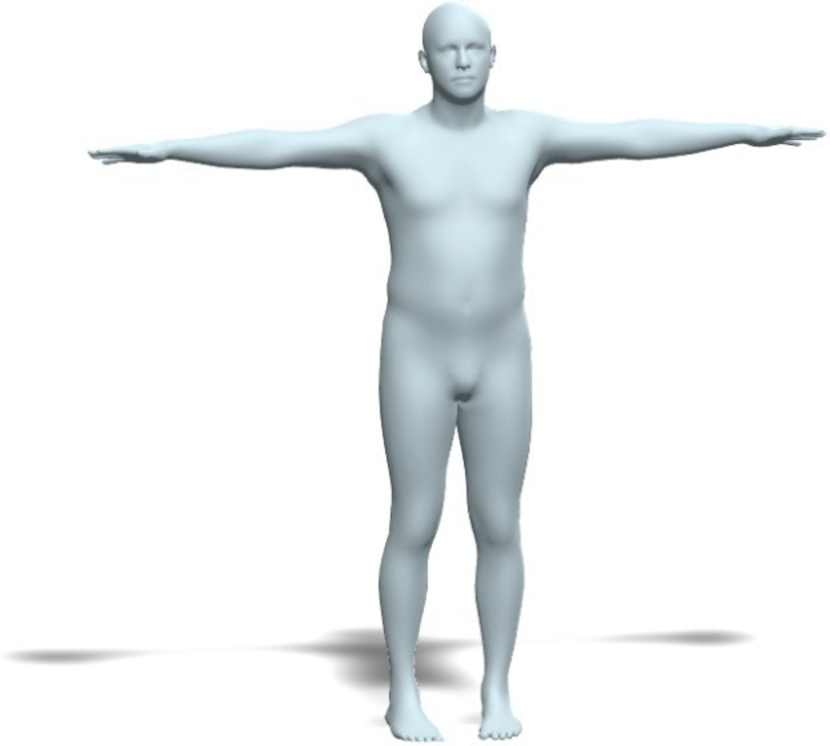


Feedforward NN

Failed Attempt II

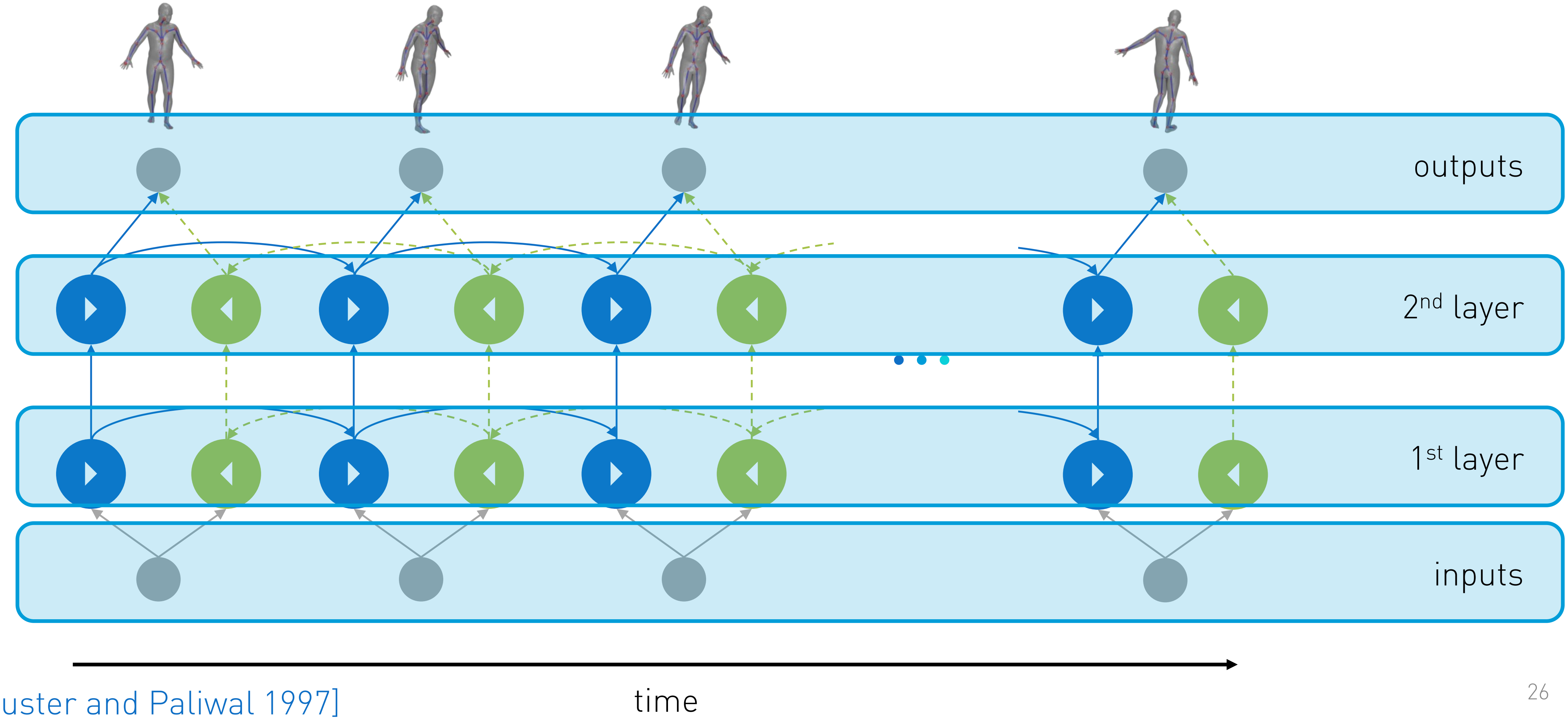


Reference



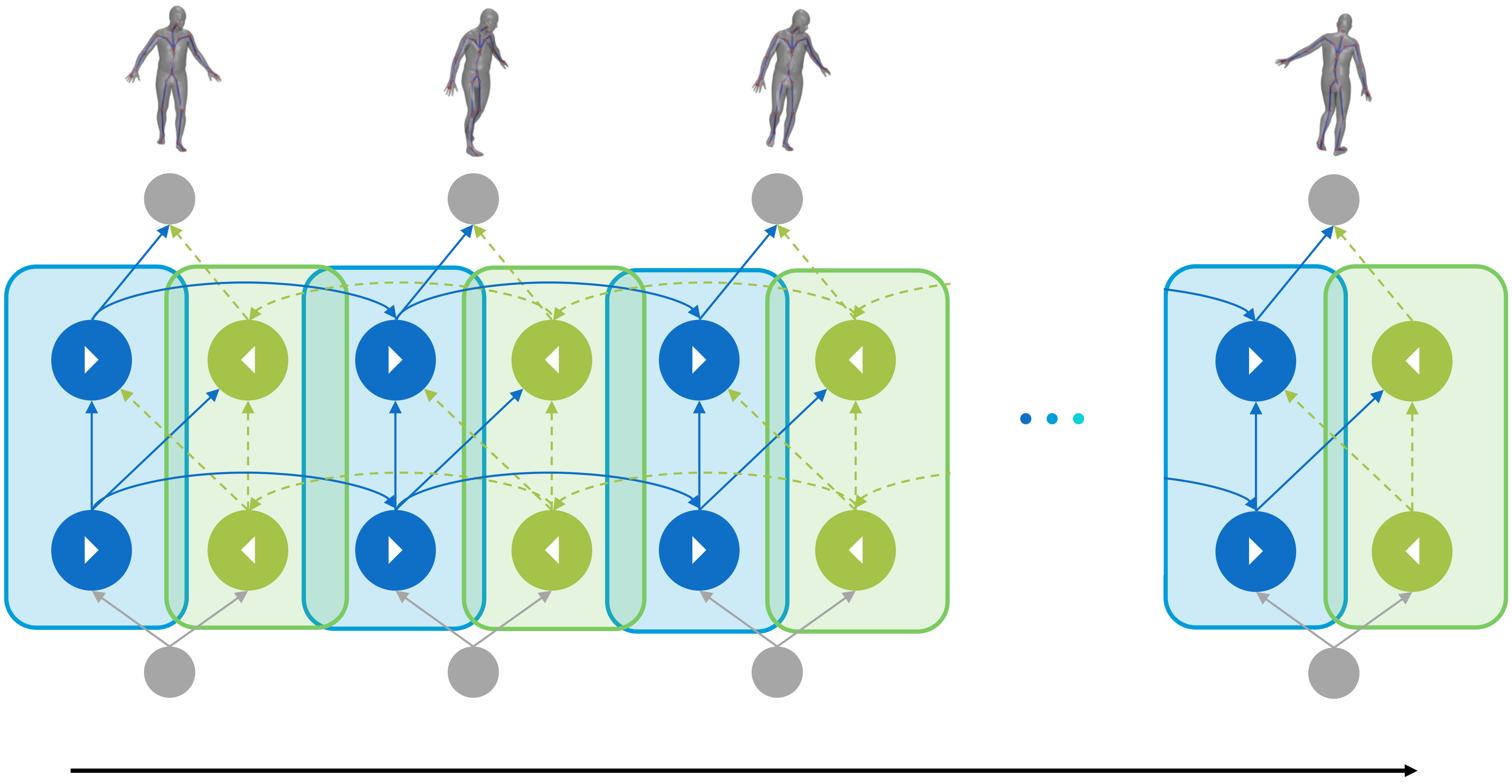
WaveNet
[van den Oord et al. 2016]

Method – Stacked BiRNN



[BiRNN: Schuster and Paliwal 1997]

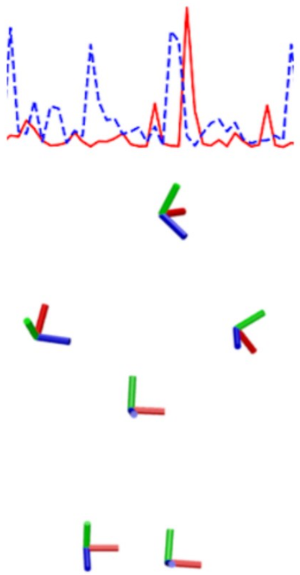
Method – Stacked BiRNN



[BiRNN: Schuster and Paliwal 1997]

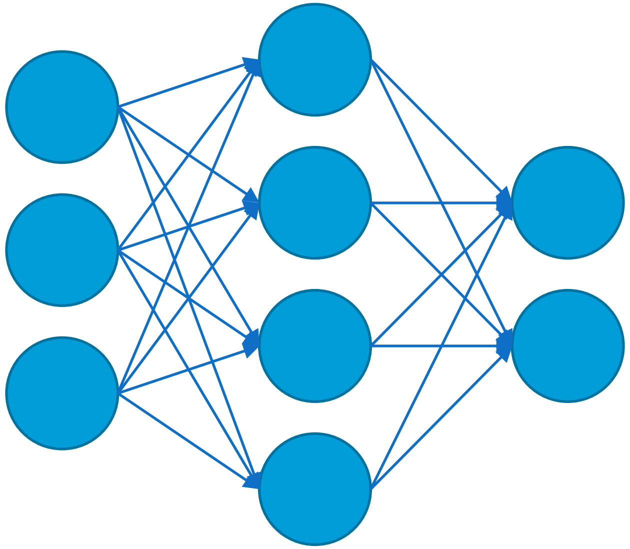
Network Design

Data

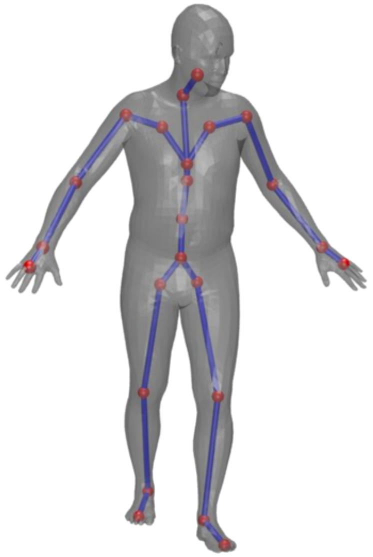


accelerations and orientations

Architecture



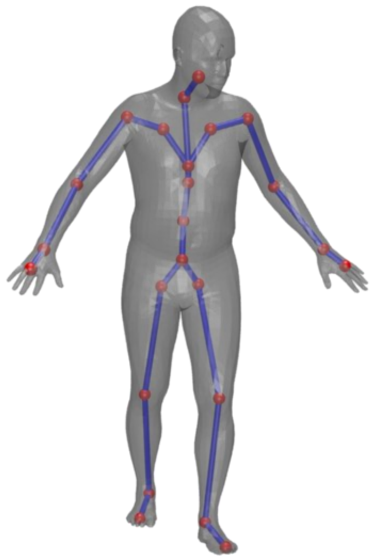
Loss Function



SMPL
pose parameters

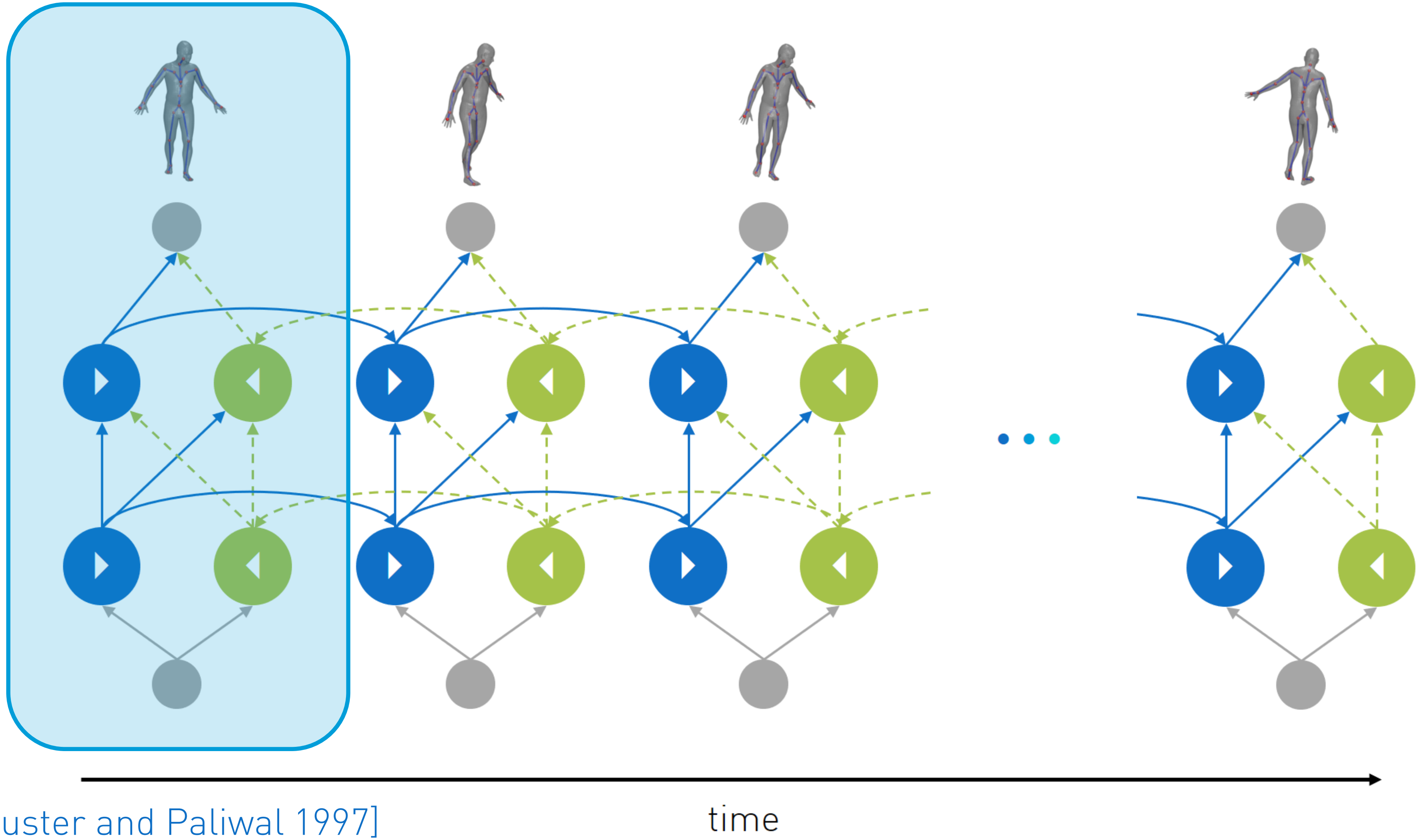
Network Design

Loss Function



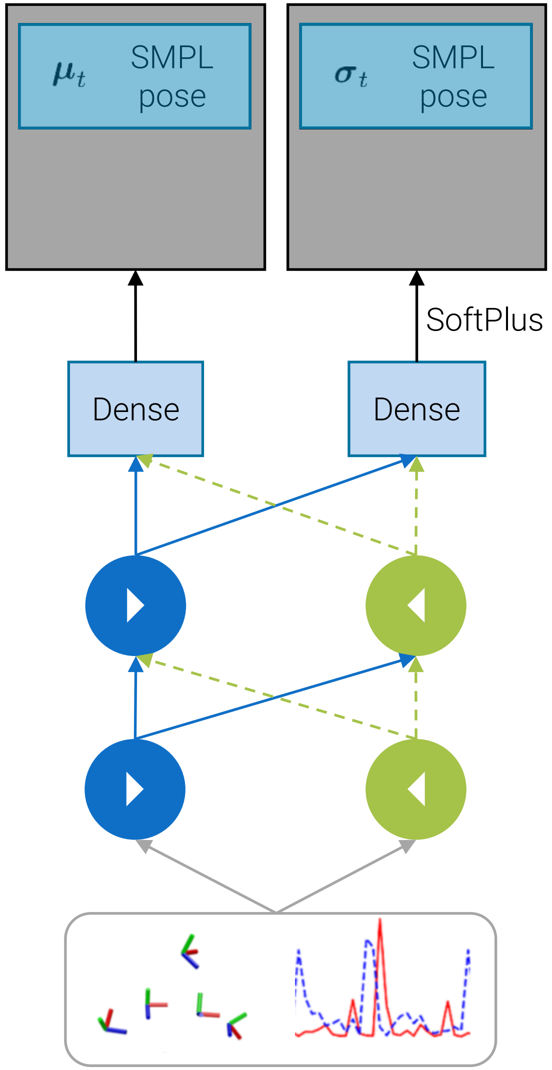
SMPL
pose parameters

Loss Function



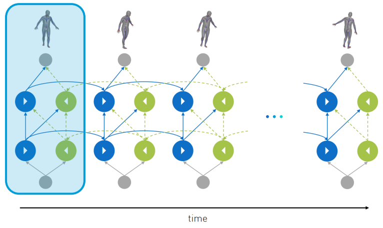
[BiRNN: Schuster and Paliwal 1997]

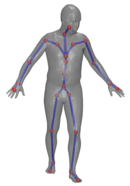
Loss Function



Pose Log-Likelihood

$$\log p(\mathbf{y}) = \sum_t^T \log \mathcal{N}(\mathbf{y}_t | \boxed{\boldsymbol{\mu}_t, \mathrm{diag}(\boldsymbol{\sigma}_t)})$$



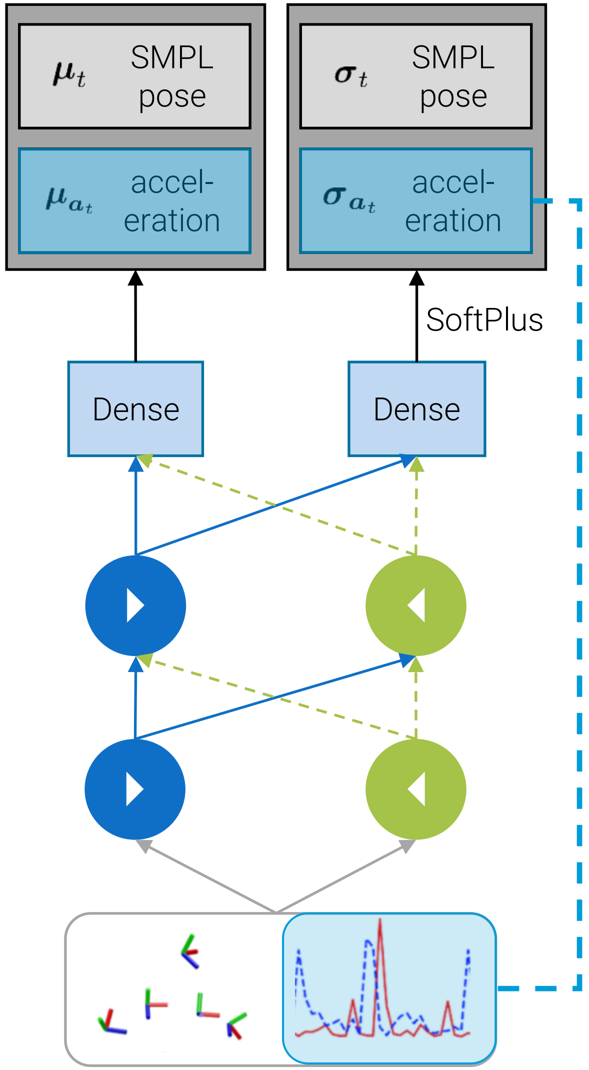


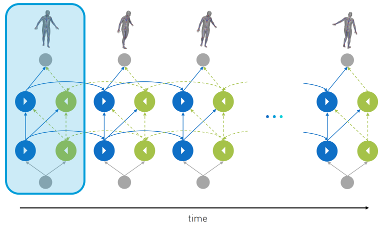
Pose, we used $\boldsymbol{\theta}$ earlier in the lecture

Question: What happens to the likelihood if the predicted variance is high?

Question: When will the network predict high variance?

Loss Function

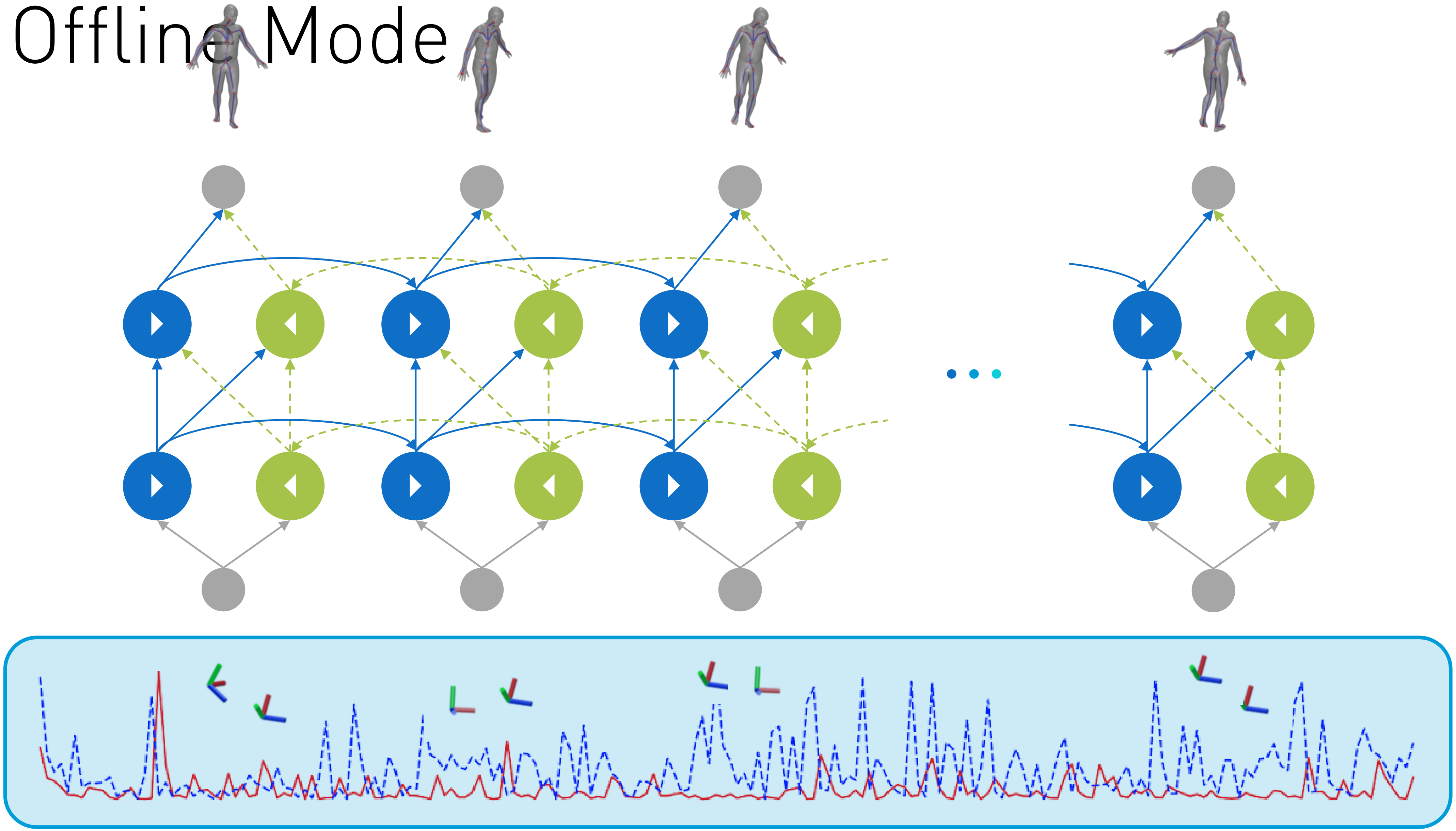




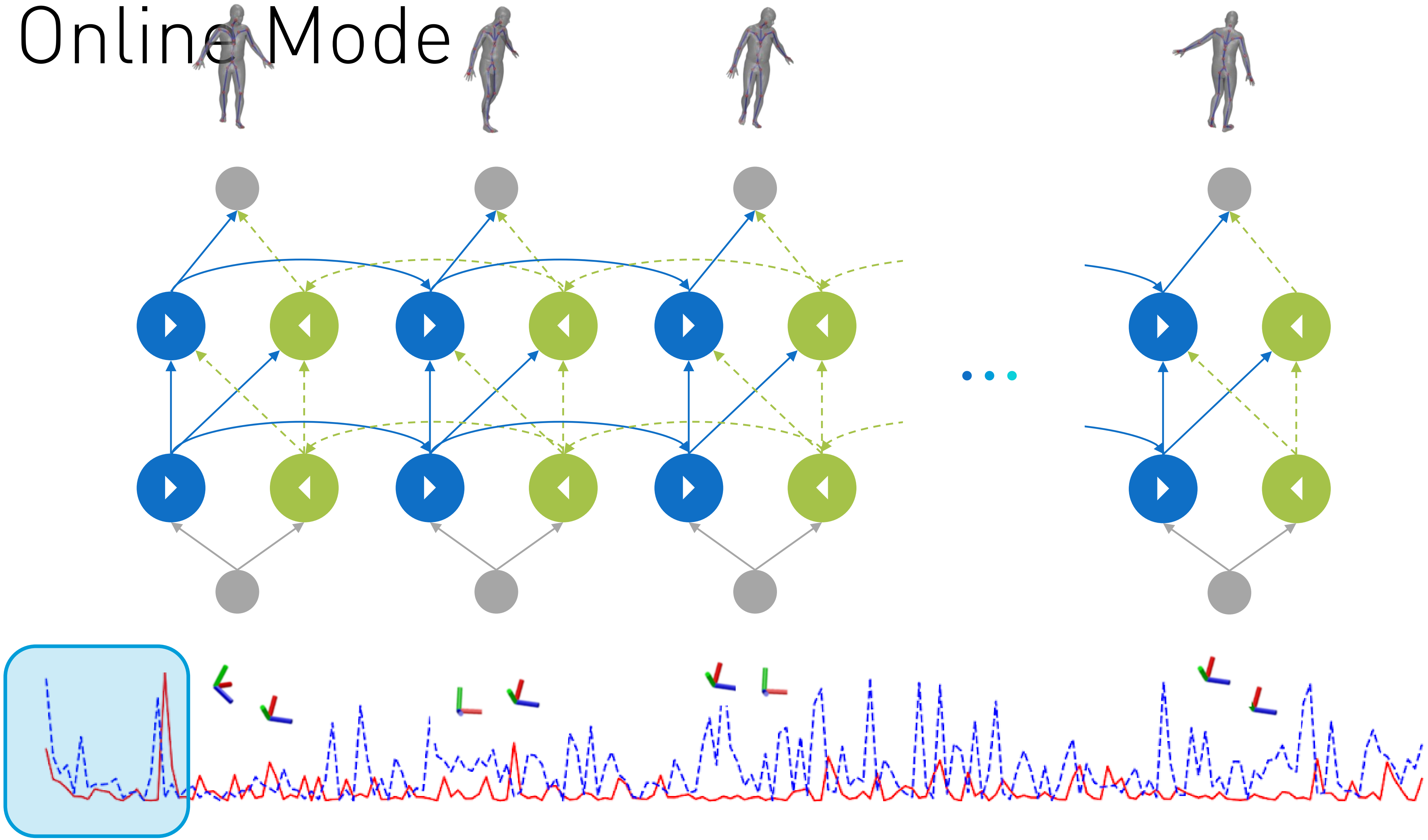
Acceleration Reconstruction Log-Likelihood

$$\log p(\mathbf{a}) = \sum_{t}^{T} \log \mathcal{N}(\mathbf{a}_t | \boxed{\boldsymbol{\mu}_{\mathbf{a}_t}, \mathrm{diag}(\boldsymbol{\sigma}_{\mathbf{a}_t})})$$

Offline Mode



Online Mode




Results

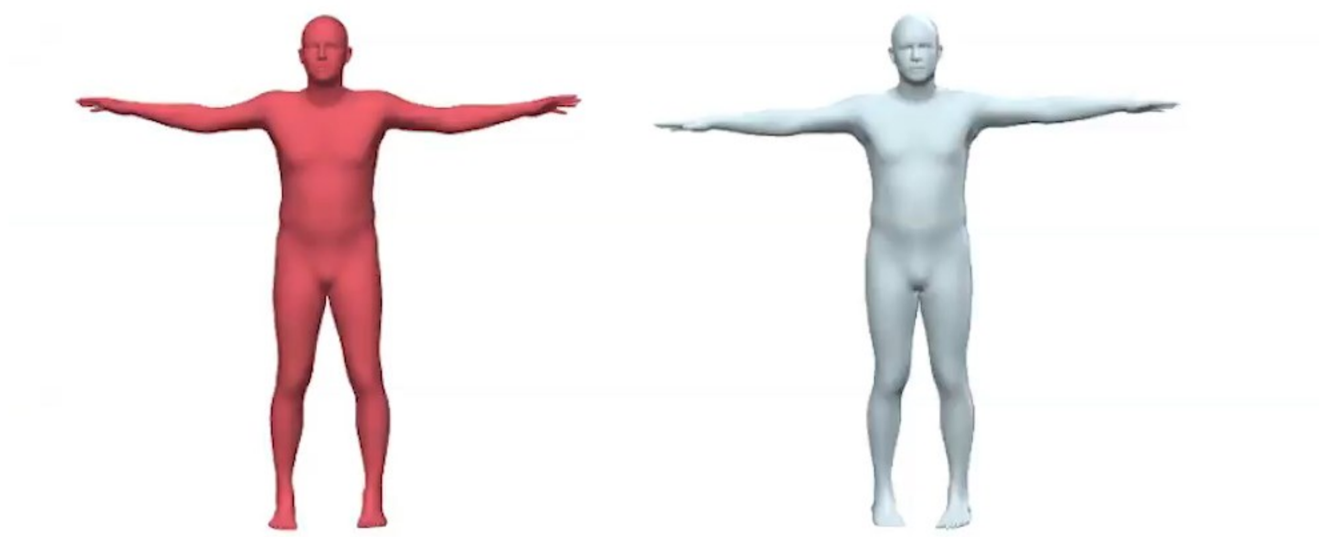
TotalCapture (offline)

[Trumble et al. 2017]

Reference SOP Ours (DIP)

TotalCapture (offline)

[Trumble et al. 2017]

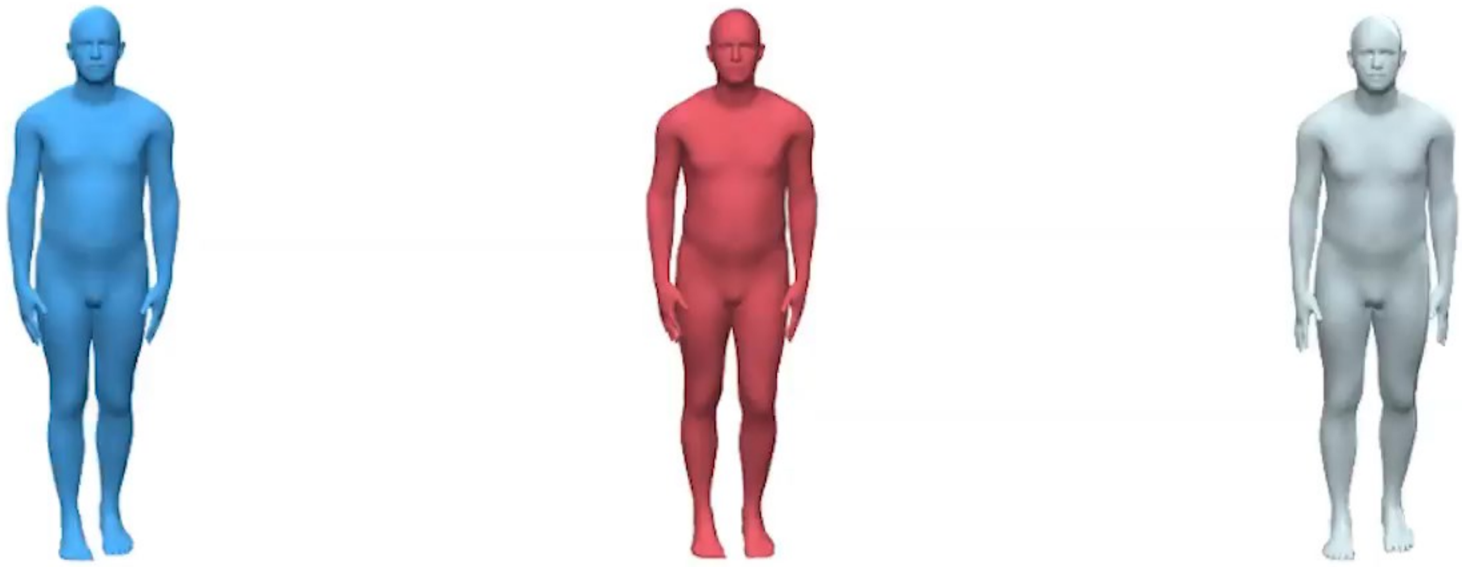


SIP

Ours (DIP)

Playground (offline)

[von Marcard et al. 2017]



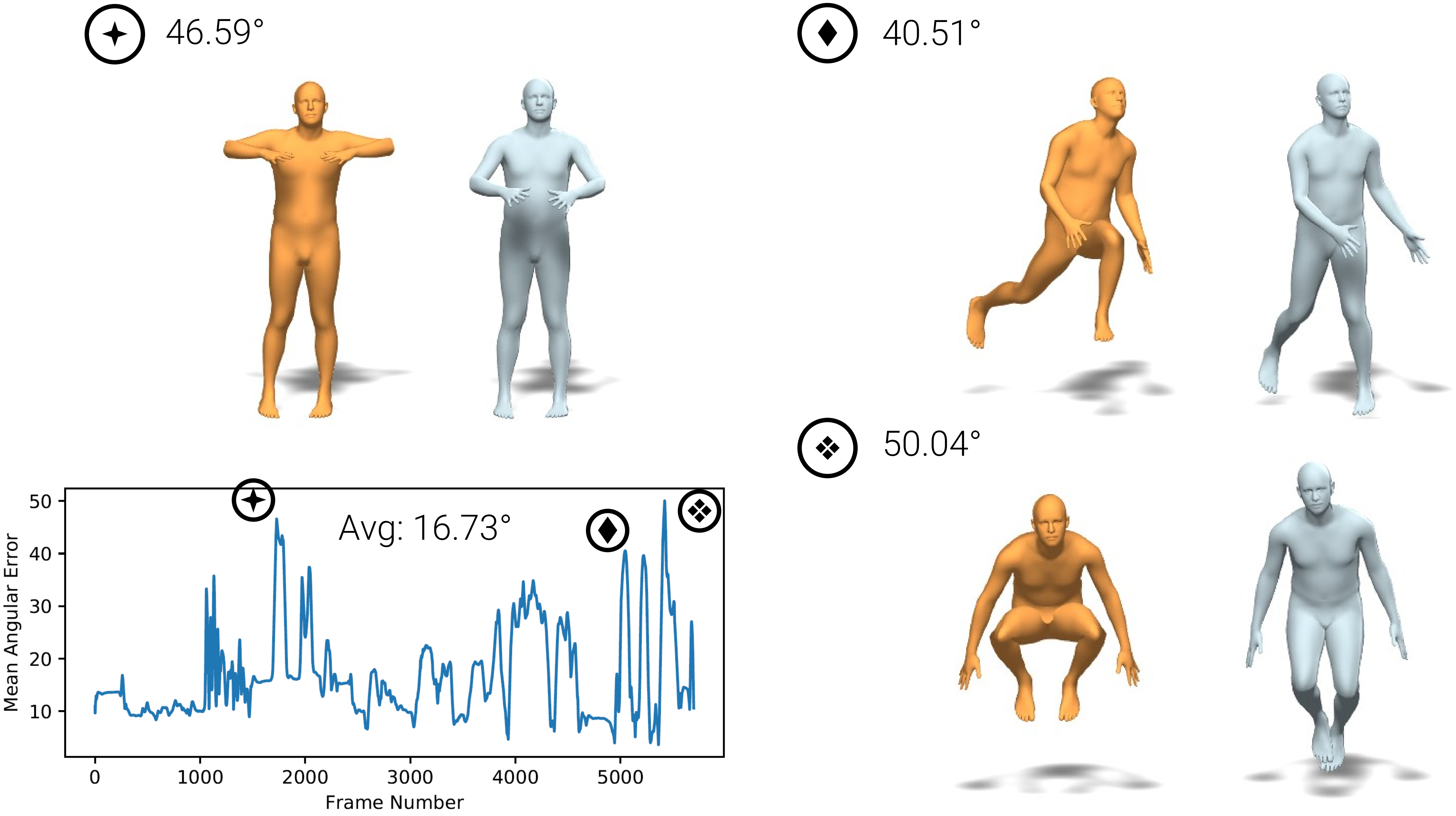
SOP SIP Ours (DIP)

Metrics on TotalCapture [Trumble et al. 2017]

	TotalCapture			
	μ_{ang}[deg]	σ_{ang}[deg]	μ_{pos}[cm]	σ_{pos}[cm]
	22.18	17.34	8.39	7.57
	16.98	13.26	**5.97**	5.50
e)	**15.85**	12.87	5.98	6.03

mean joint angle error

mean positional error



46.59°
40.51°
50.04°
Avg: 16.73°
Mean Angular Error
Frame Number
50
40
30
20
10
0
1000
2000
3000
4000
5000

Real-Time Performance

System should work with **real** data in **real-time**.

Not a given as **noise characteristics** might by very different.

DIP-IMU Dataset

Recorded our own **dataset** with **17 Xsens sensors.**

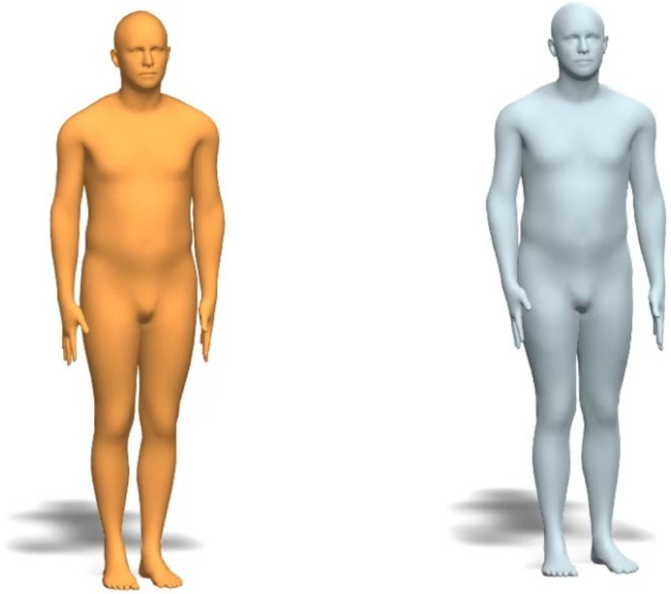
Feed **SIP** fully-constrained pose to produce reference SMPL poses **(SIP-17)**.

10 subjects, roughly **90 min.** of data.

http://dip.is.tue.mpg.de

Fine-Tuning for Domain Adaptation

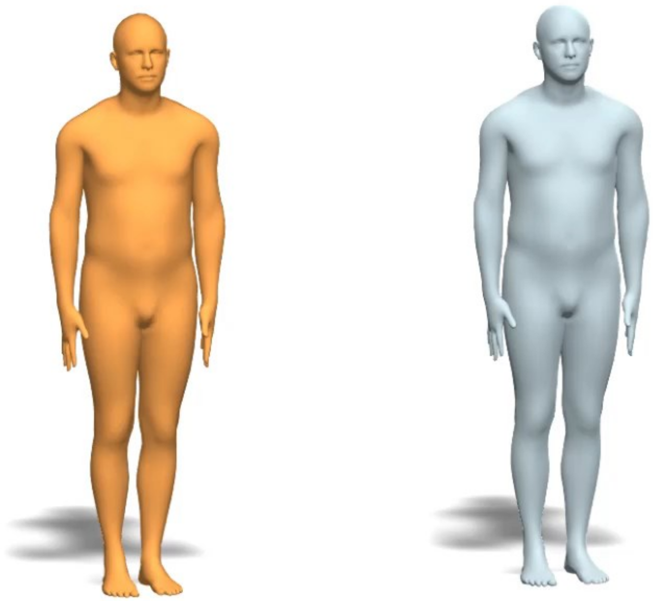
Domain adaptation problem **severe** on DIP-IMU.



Reference (SIP-17)

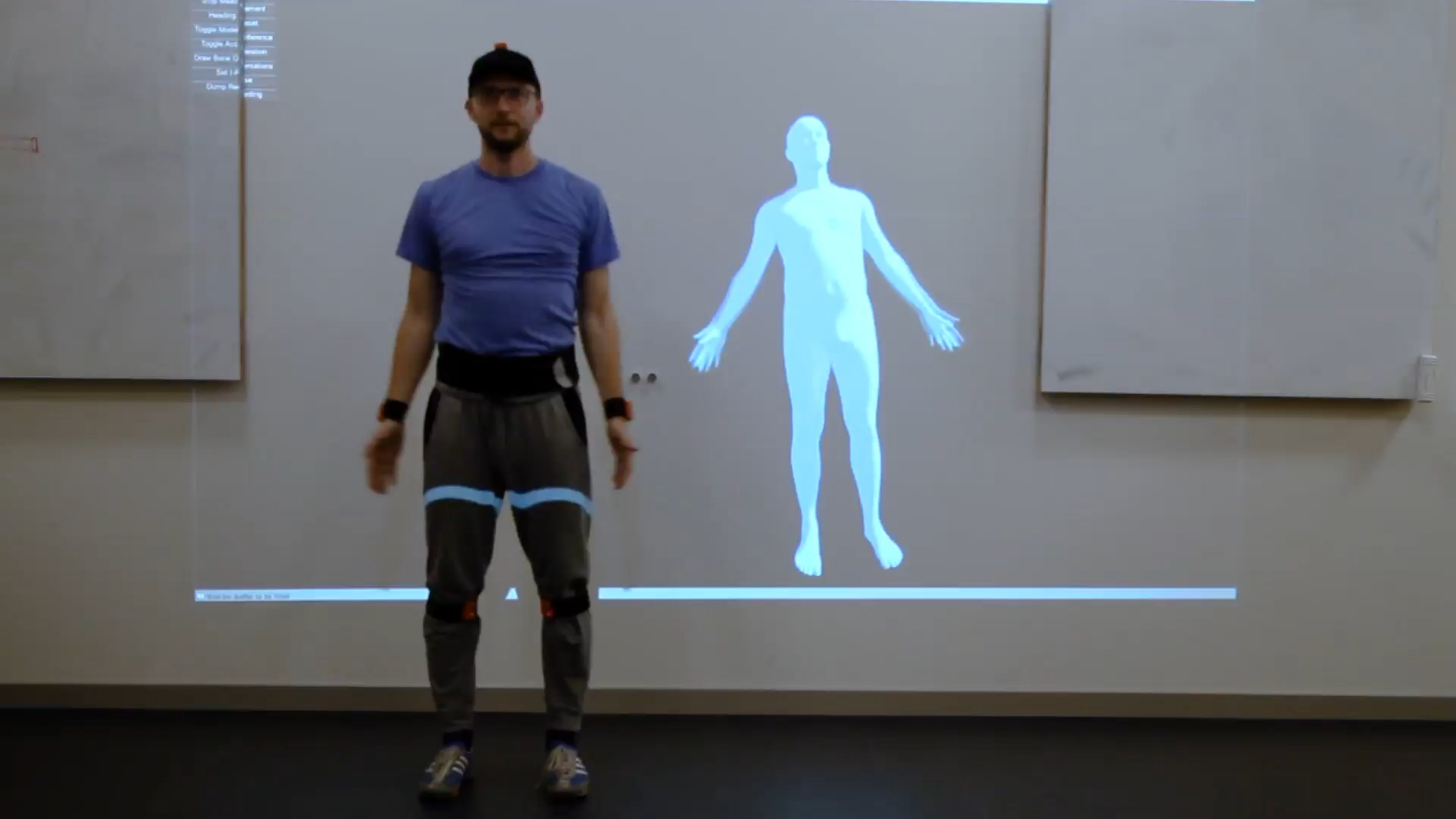
Ours (before fine-tuning)

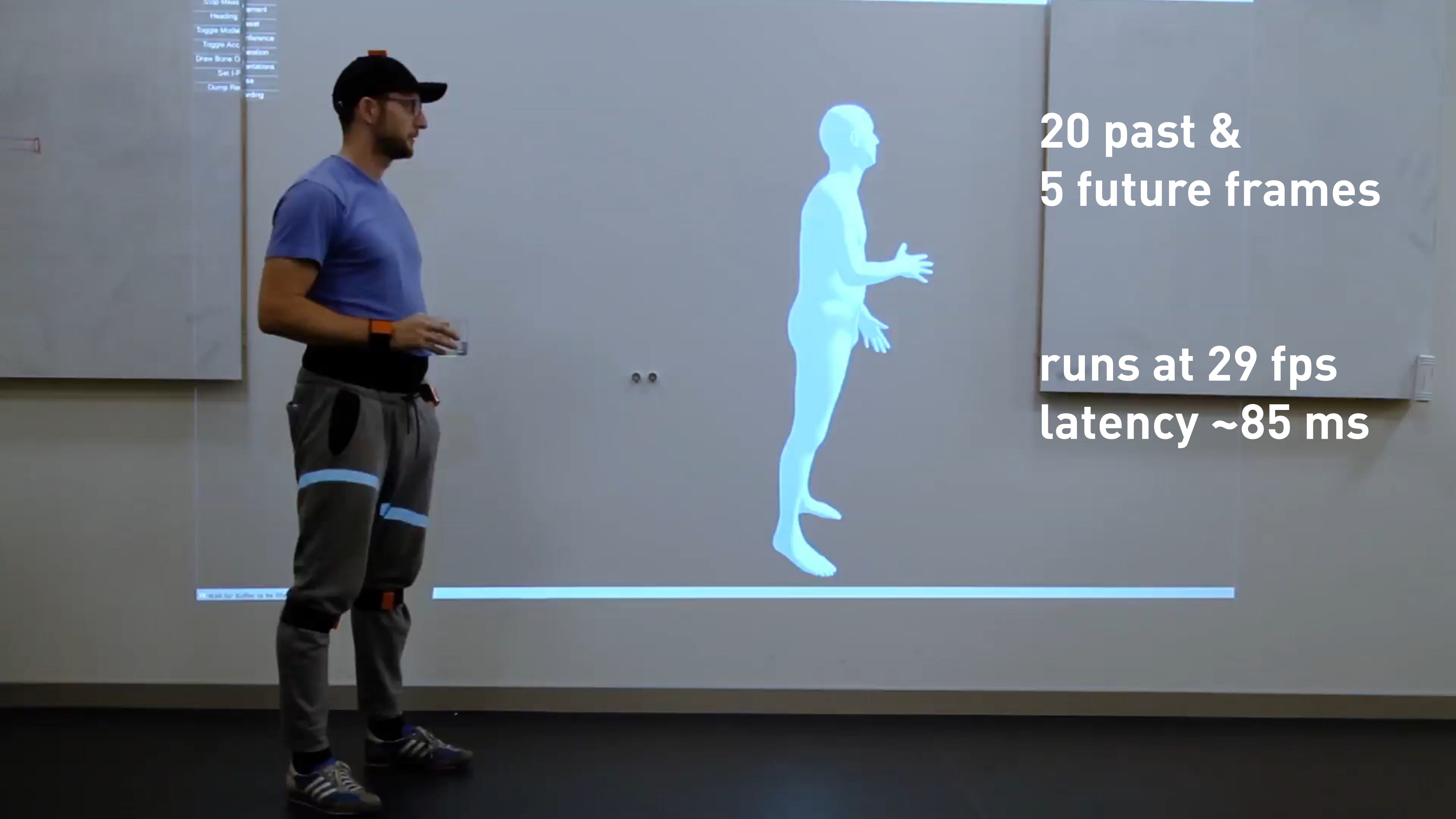
After **fine-tuning** on subset of DIP-IMU.

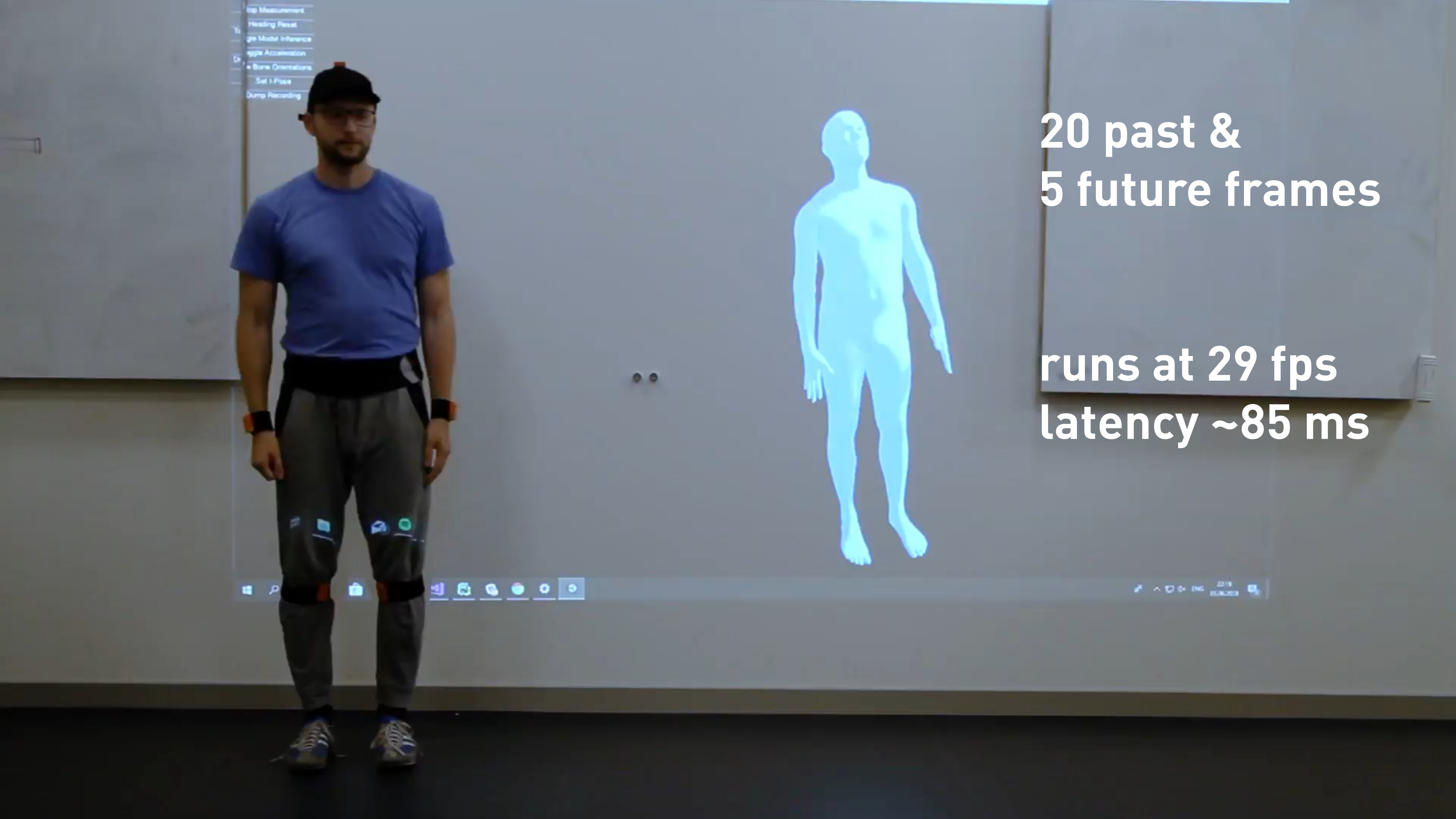


Reference (SIP-17)

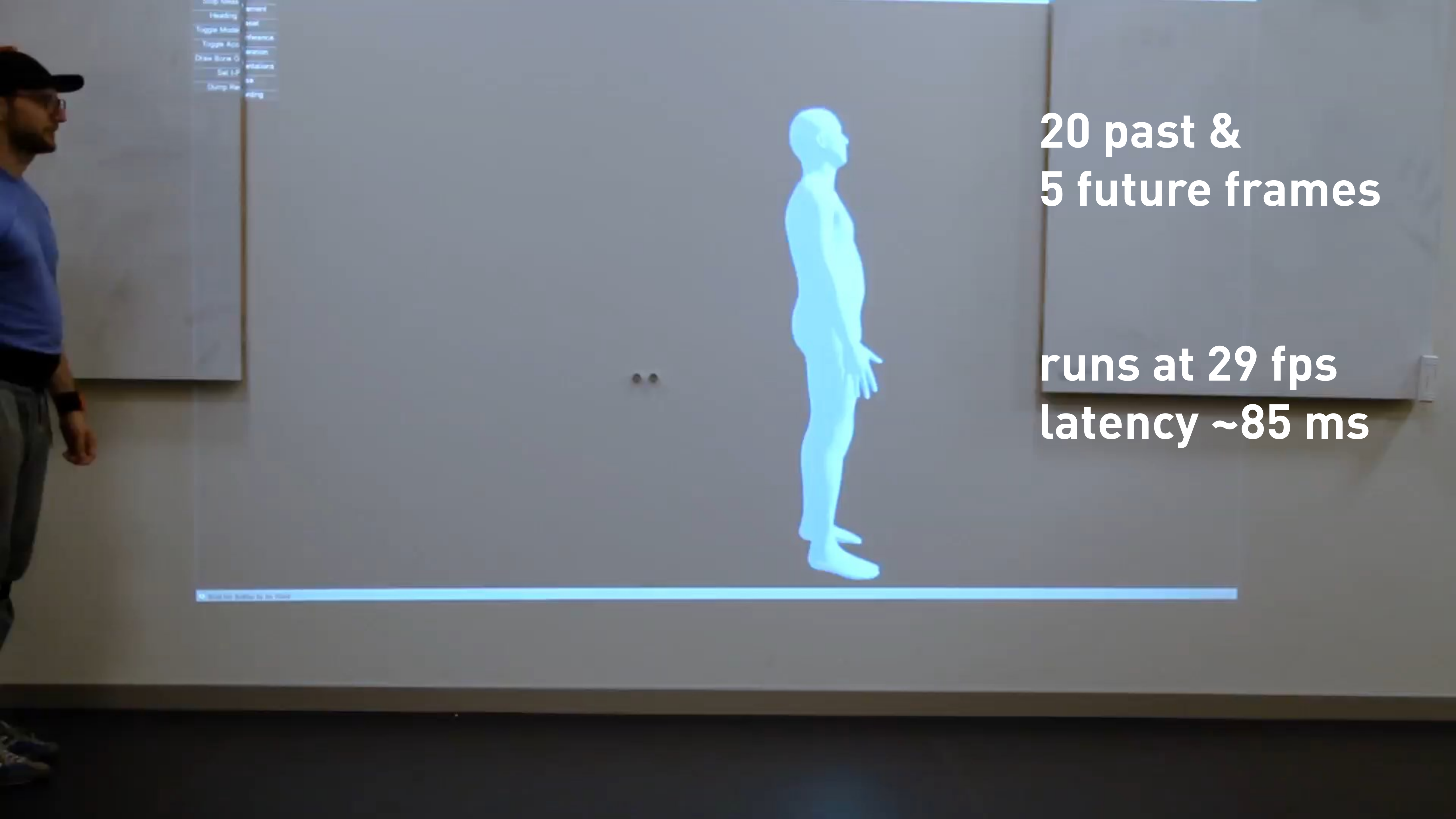
Ours (after fine-tuning)




20 past &
5 future frames
runs at 29 fps
latency ~85 ms



20 past &
5 future frames
runs at 29 fps
latency ~85 ms


20 past &
5 future frames
runs at 29 fps
latency ~85 ms

Summary

Possible to capture motions in **real time** with **sparse** set of IMUs.

Training on large **synthetic** dataset.

Domain adaptation still difficult.

We **release** code and data.

http://dip.is.tue.mpg.de

Thank You!

Deep **I**nertial **P**oser

Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time

http://dip.is.tue.mpg.de

References

Ahmed **Elhayek**, Edilson de Aguiar, Arjun Jain, J Thompson, Leonid Pishchulin, Mykhaylo Andriluka, Christoph Bregler, Bernt Schiele, and Christian Theobalt. 2017. MARCOnI ConvNet-Based MARker-Less Motion Capture in Outdoor and Indoor Scenes. IEEE transactions on pattern analysis and machine intelligence 39, 3 (2017), 501–514.

Matthew **Loper**, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) 34, 6 (2015), 248.

Dushyant **Mehta**, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. 2018. Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB. In International Conference on 3D Vision (3DV)

Charles **Malleson**, Marco Volino, Andrew Gilbert, Matthew Trumble, John Collomosse, and Adrian Hilton. 2017. Real-time Full-Body Motion Capture from Video and IMUs. In 2017 Fifth International Conference on 3D Vision (3DV). 449–457.

Daniel **Roetenberg**, Henk Luinge, and Per Slycke. 2007. Moven: Full 6dof human motion tracking using miniature inertial sensors. Xsen Technologies, December (2007).

Mike **Schuster** and Kuldip K **Paliwal**. 1997. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45, 11 (1997), 2673–2681.

Matthew **Trumble**, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. 2017. Total capture: 3d human pose estimation fusing video and inertial sensors. In Proceedings of 28th British Machine Vision Conference. 1–13.

Timo **von Marcard**, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. 2017. Sparse inertial poser: Automatic 3D human pose estimation from sparse IMUs. In Computer Graphics Forum, Vol. 36. Wiley Online Library, 349–360.

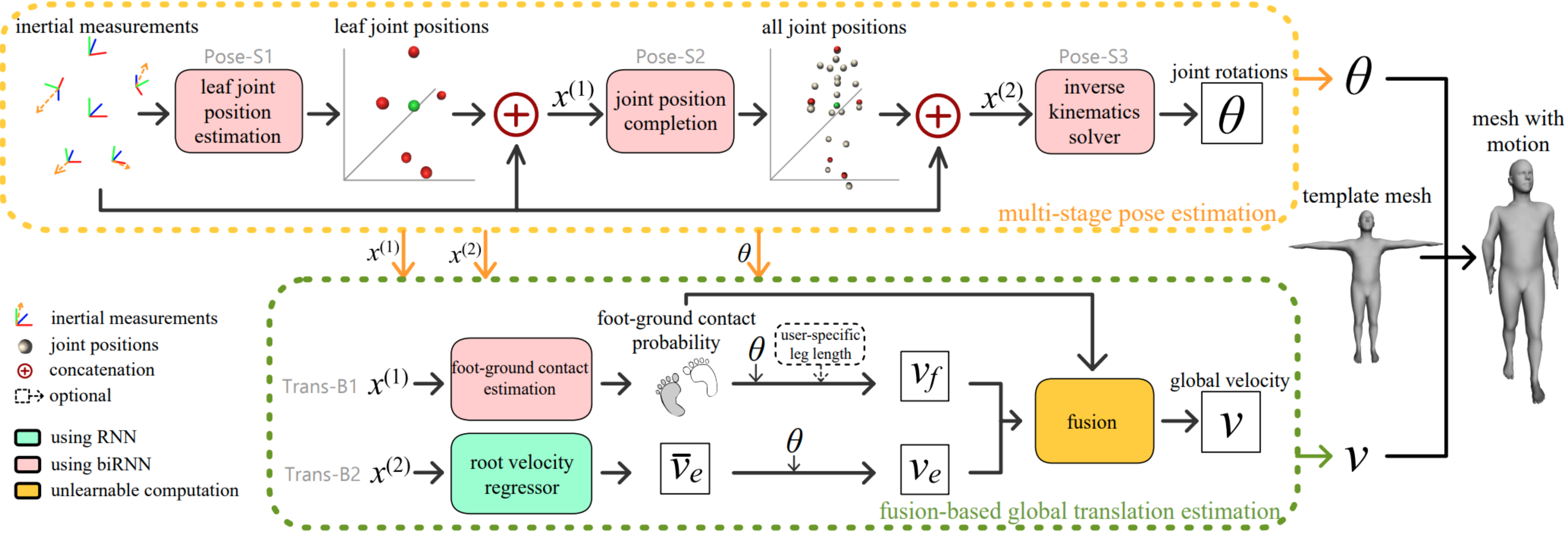
Timo **von Marcard**, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In European Conference on Computer Vision (ECCV)

Extensions and recent works

TransPose: Global translation and physical constraints

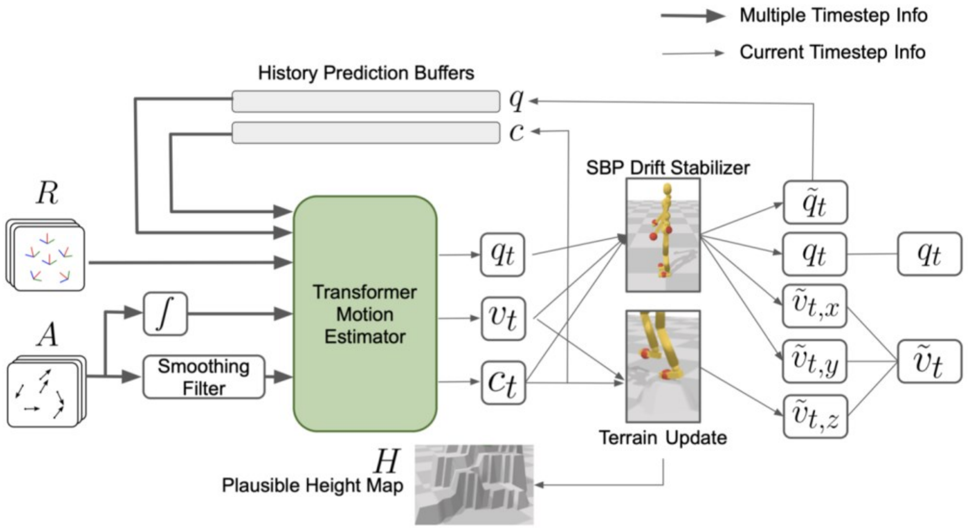
Key ideas to improve DIP:

1) Predict joints from leaf to root hierarchically
2) Predict and enforce foot contact to the ground

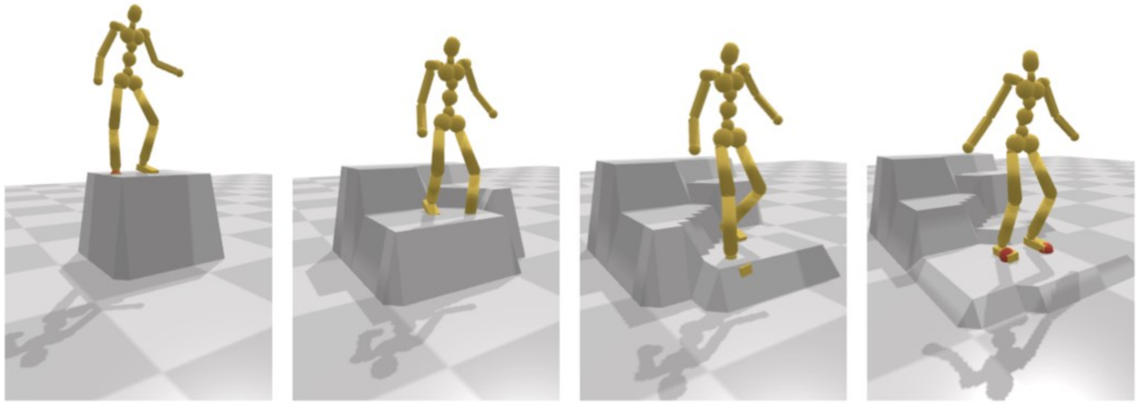


Yi et al. Siggraph'21

Trasformer Inertial Poser



Architecture



Example of predicted terrain

Key ideas:

1) Predict stationary points to constraint motion
2) Infer plausible terrain

Jiang et al. SA’22

PIP: Physical Inertial Poser

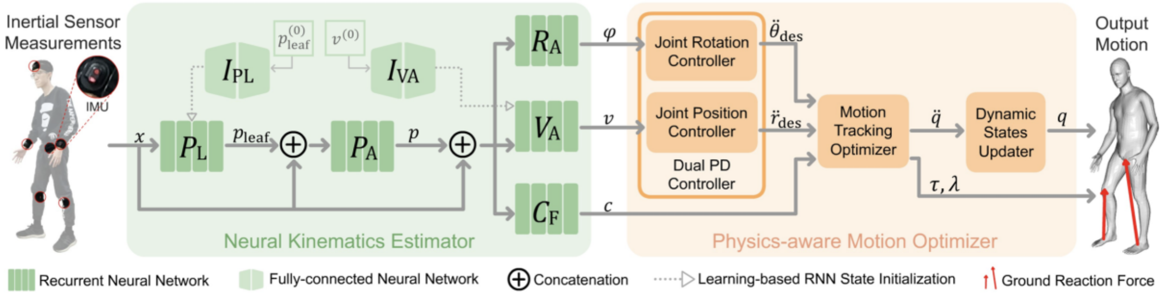
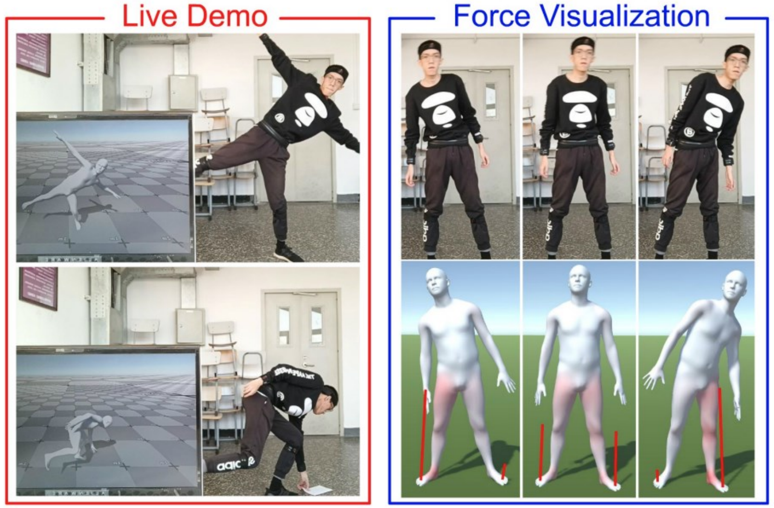


Figure 2. Overview of our method. We first use a neural kinematics estimator to infer human motion status from sparse IMU measurements. Then, we use a physics-aware motion optimizer to obtain physically correct human motion, joint torques, and ground reaction forces.



Key idea:

1) Predict Motion with a neural model
2) Refine estimate with physics based optimization (need to figure external forces as well as body joint torques)

Yi et al. CVPR’22

Slide Acknowledgments

Manuel Kaufmann, Yinghao Huang