Perceiving Humans

Angjoo Kanazawa

CS280

March 31, 2025

Logistics

- Today: 2D/3D Humans
- HW3 up on keypoint detection
- Wednesday: Jitendra
- Next Monday: Learning to predict correspondences
  - → Released papers for you to read in advance on Ed
- Today after class project proposal

Perceiving Humans

From Recognition to Detection to Reconstruction

Why perceive humans?

- Well they are the most important thing

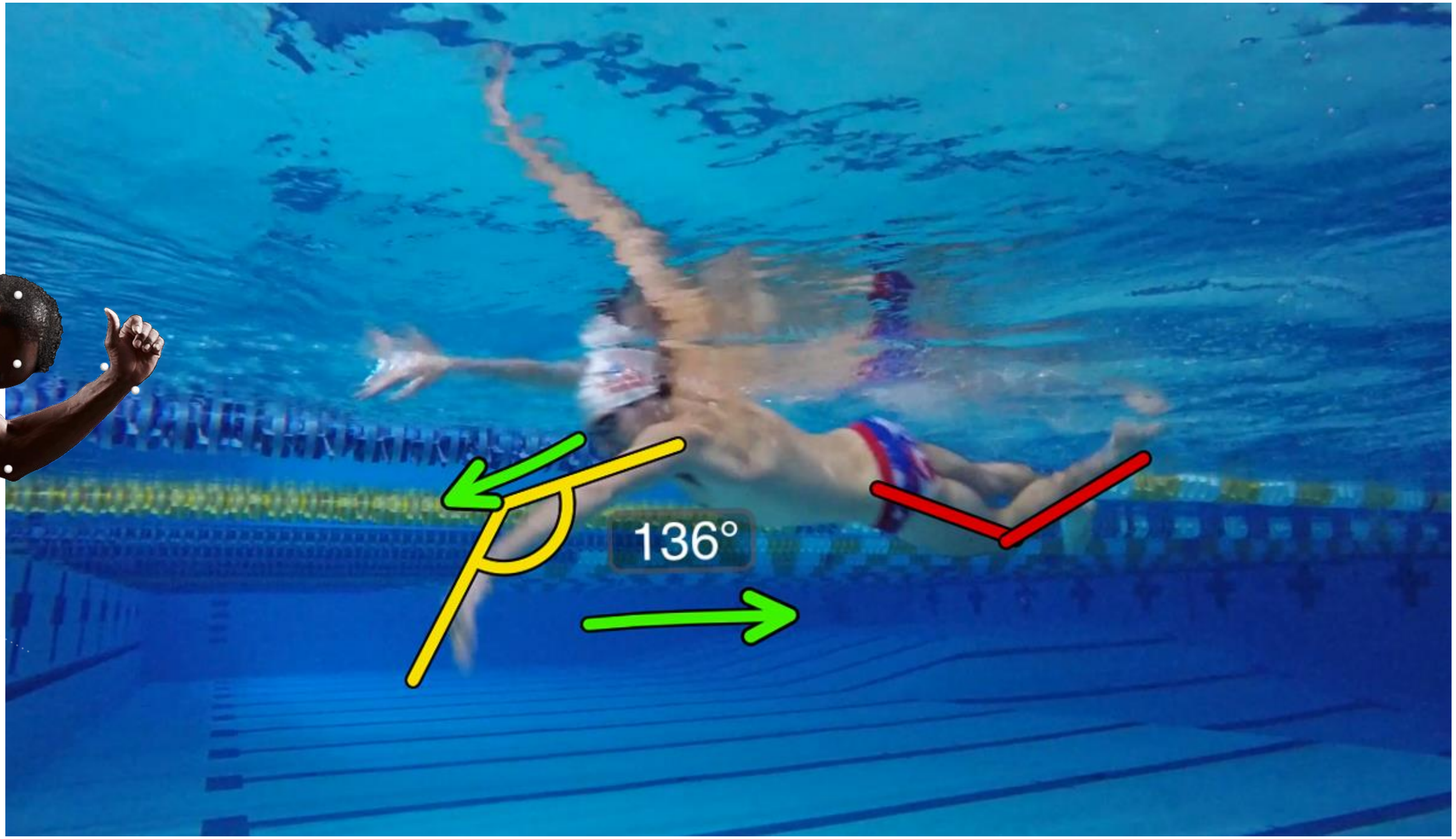



Learning to act from visual observation

Anticipating human behavior

Sport analysis

OptiTrack



MySwimPro

Medical diagnosis and treatment

Photo Credit: Qualisys

Challenges

Why is perceiving humans hard?


variation in appearance

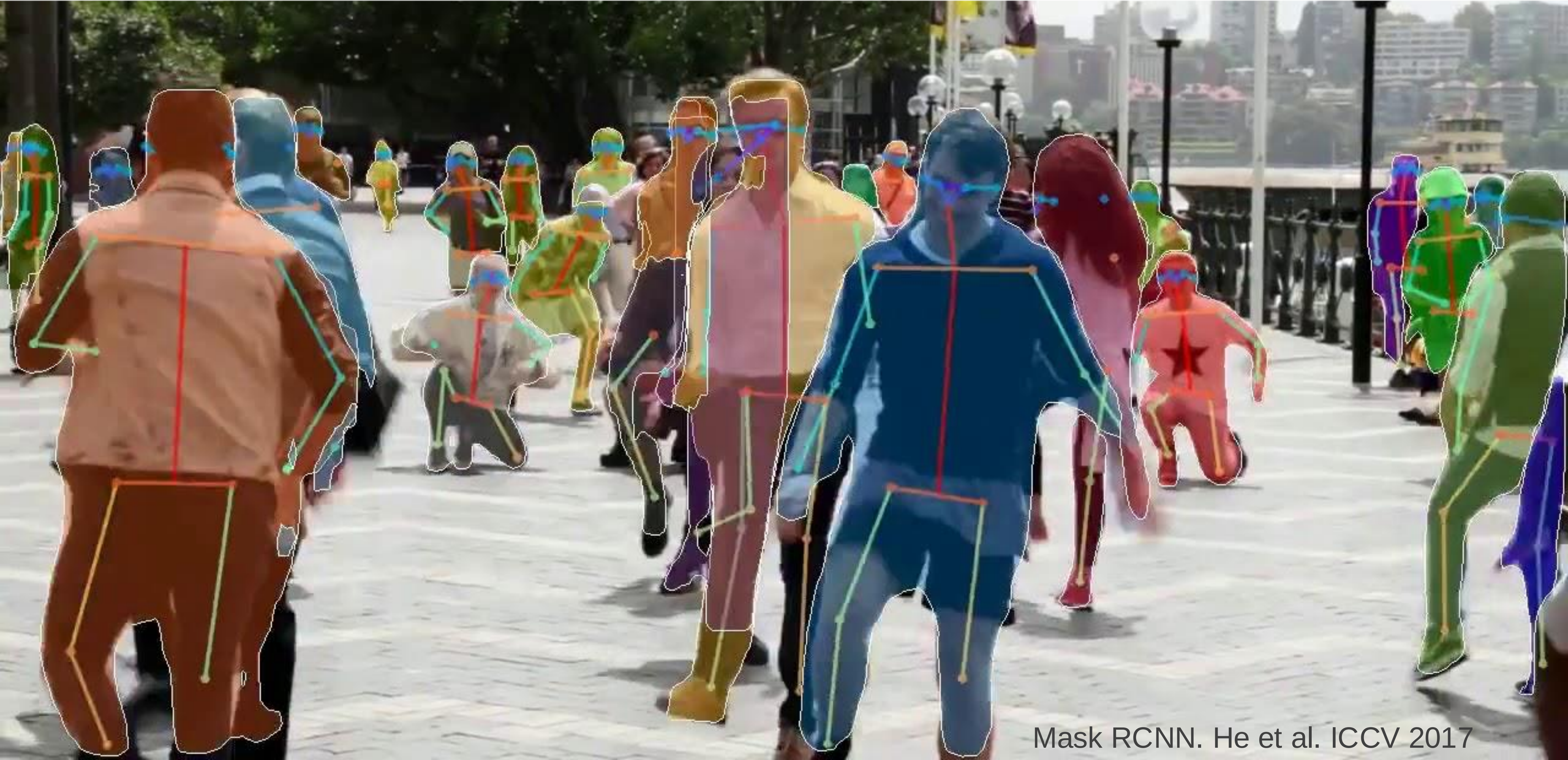
variation in pose, viewpoint


occlusion & clutter


variation in illumination

Slide Credit: Deva Ramanan

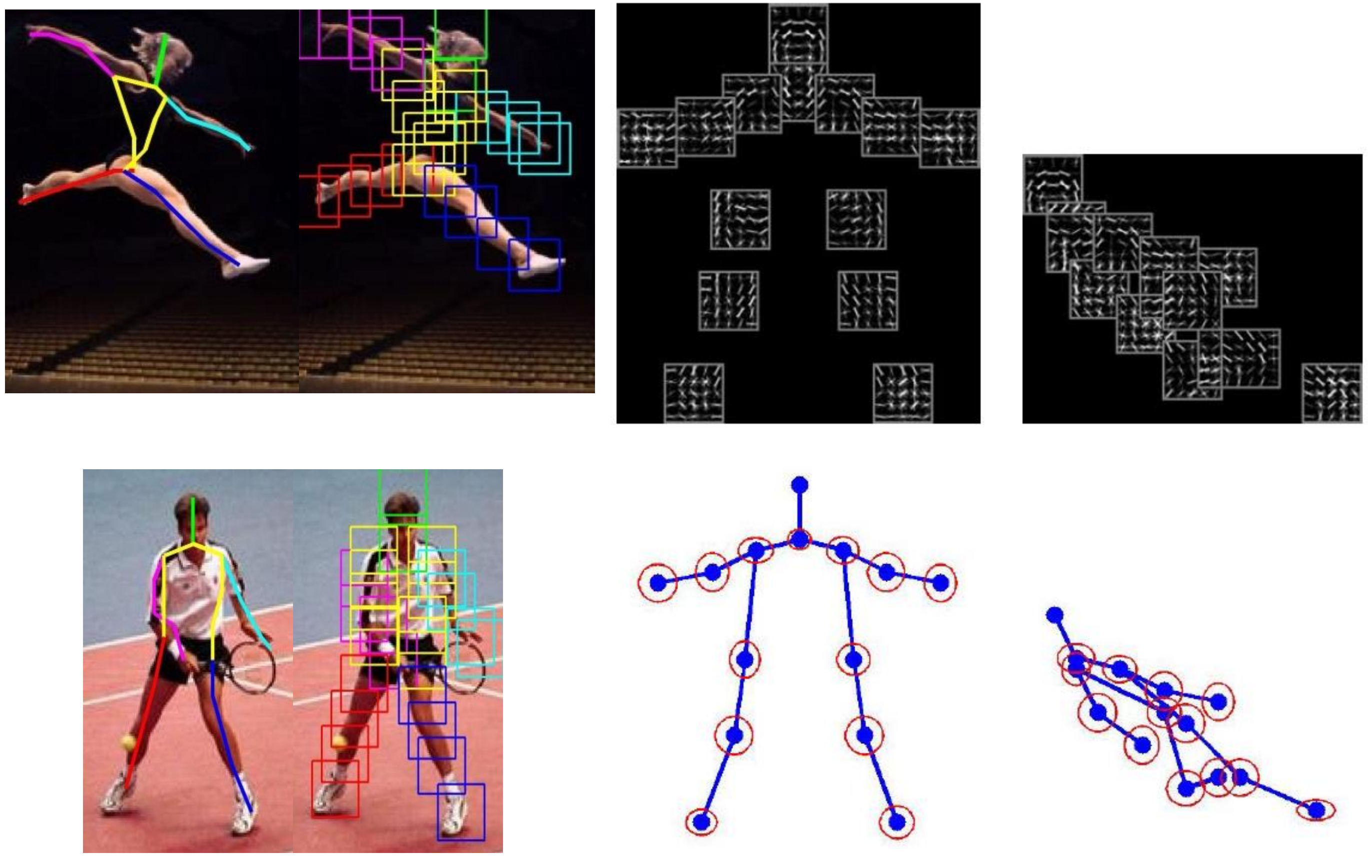
2D Humans



Mask RCNN. He et al. ICCV 2017

Parts develop finer into joints & keypoints

[Ferrari, Marín-Jiménez and Zisserman CVPR ‘08]



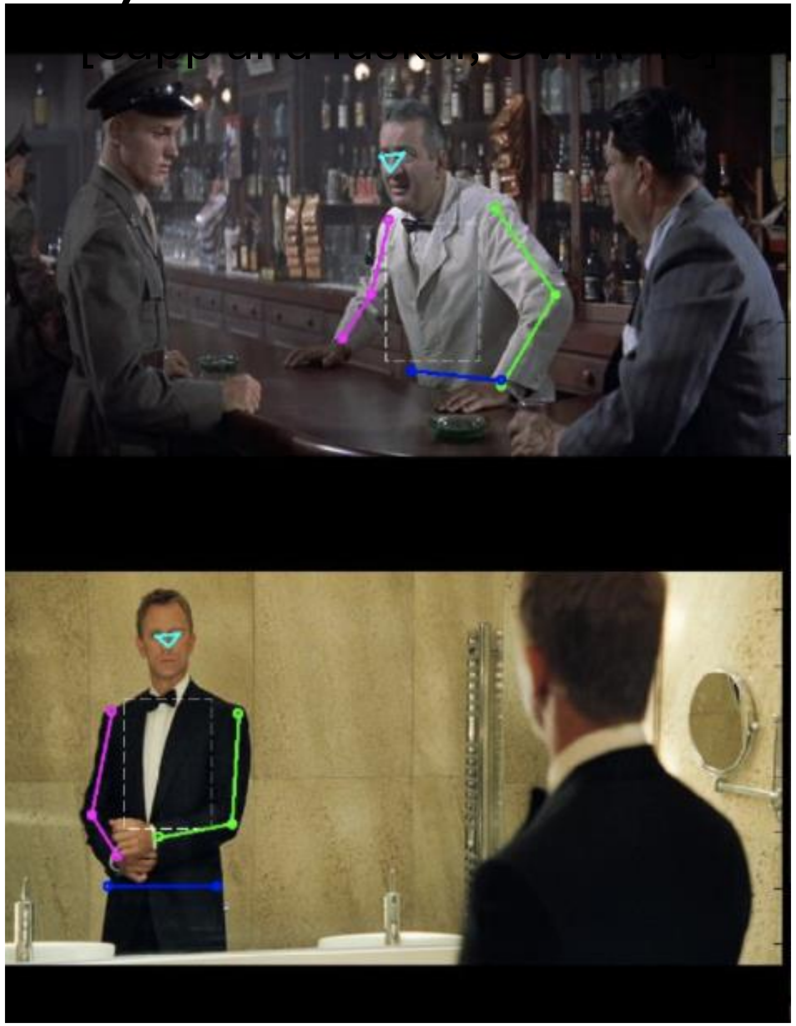
Articulated Human Pose Estimation with Flexible Mixtures of Parts
[Yang and Ramanan CVPR ‘11]

Datasets are introduced

Leeds Sports Pose (**LSP)**
[Johnson and Everingham, CVPR ‘11]

11000 Train
1000 Test

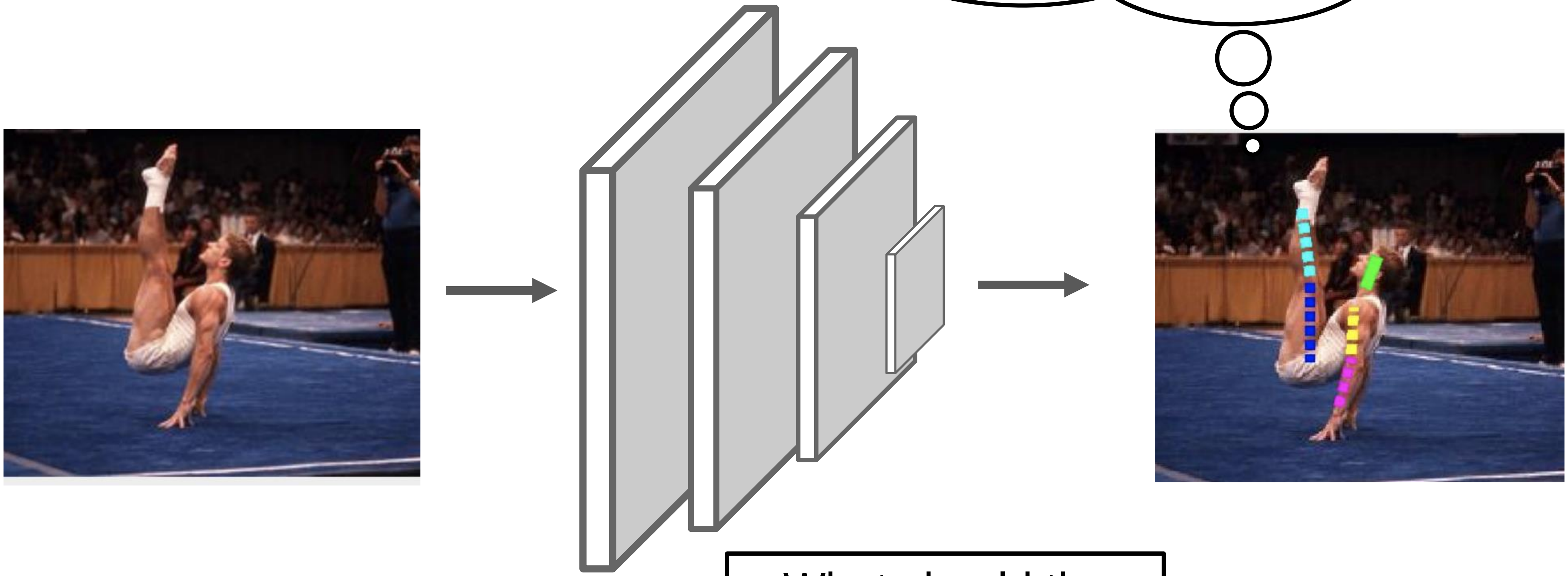
Frames Labeled in Cinema
(FLIC)



4000 Train
1000 Test

Deep Learning Era

How do we
represent/parametrize
2D human pose???

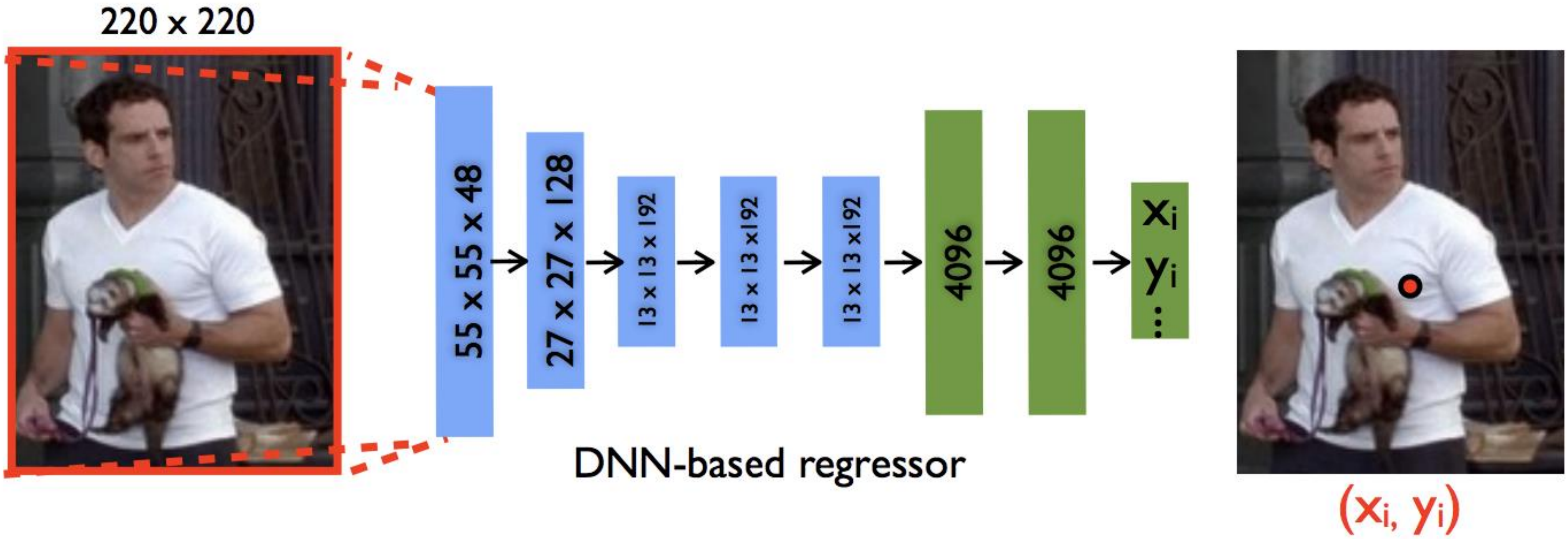


Input

What should the
network output?

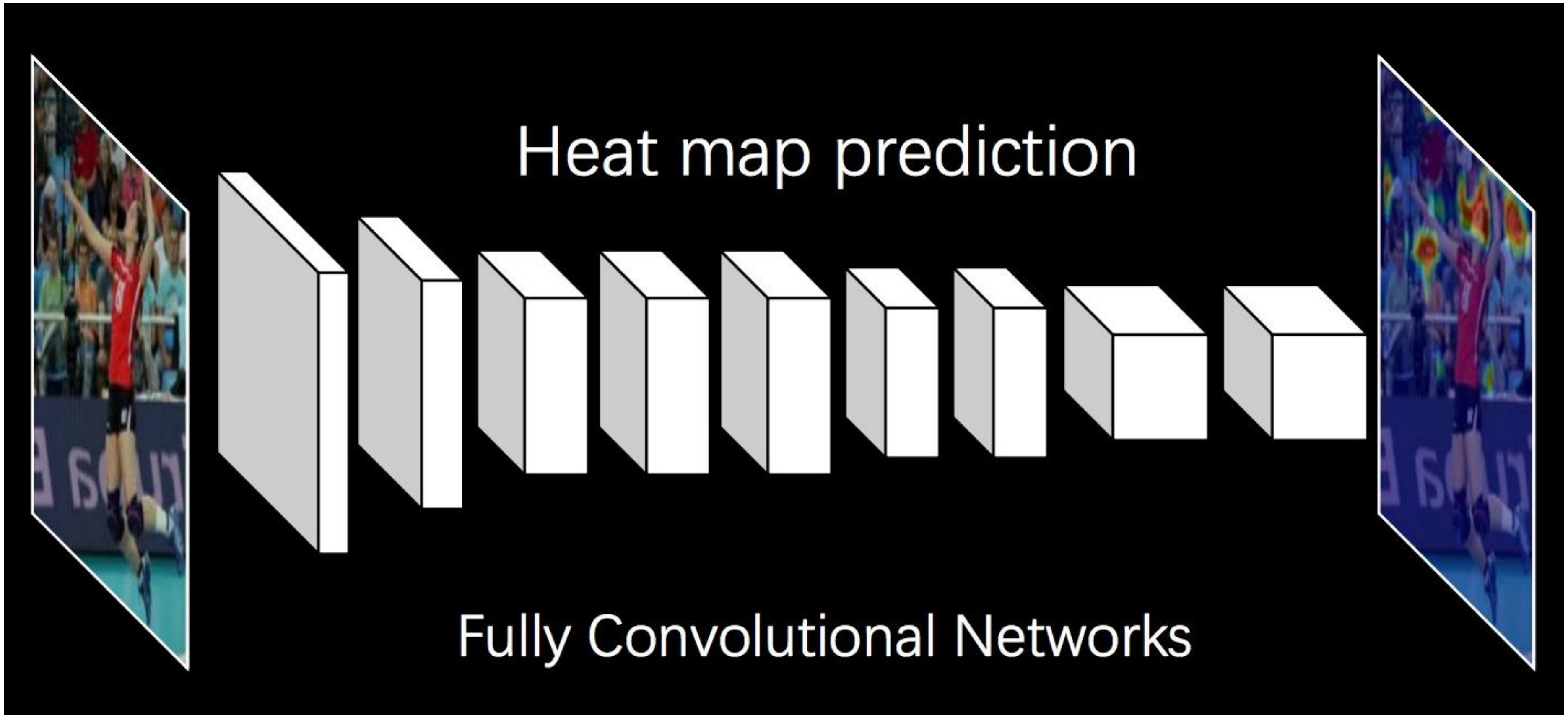
Output

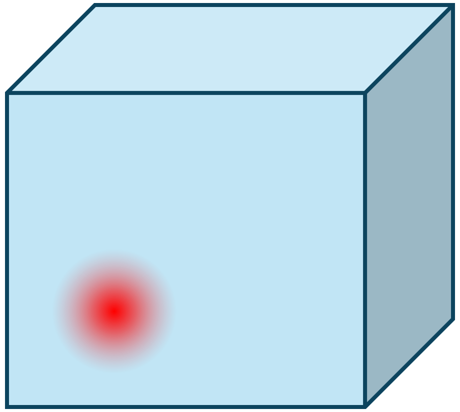
Predicting keypoints



DeepPose: Human Pose Estimation via Deep Neural Networks
[Toshev and Szegedy 2014]

Predict heat maps





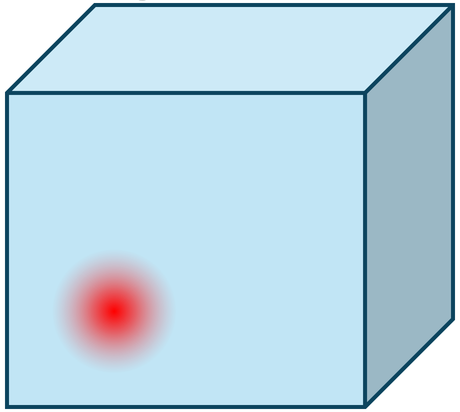
Target: K+1 x H x W
Gaussian around
(x,y) for k-th
keypoint in the k-th
channel

K+1 for K parts + background

L2 Training Loss

- L2 loss on the target heatmap (peaky gaussian around the gt keypoint)

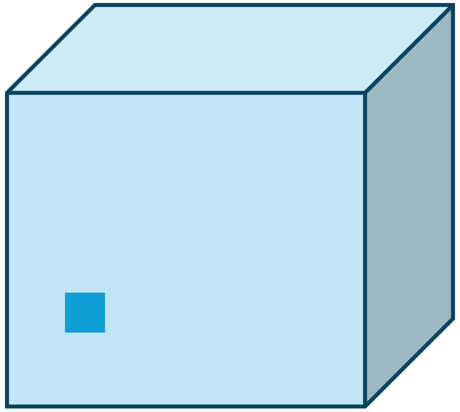
$$L = \sum_{k=1}^{K+1} \sum_{(x,y)} ||b^k(x,y) - b_*^k(x,y)||$$



Target “belief map” :
K+1 x H x W
Gaussian around (x,y) for k-th keypoint in the k-th channel

Log Loss Training Loss

- Log loss (or cross entropy loss) on the target heatmap probabilities
- The target must also sum to 1
- Mask RCNN just uses 1 at the target, 0 everywhere else.
- Experiment

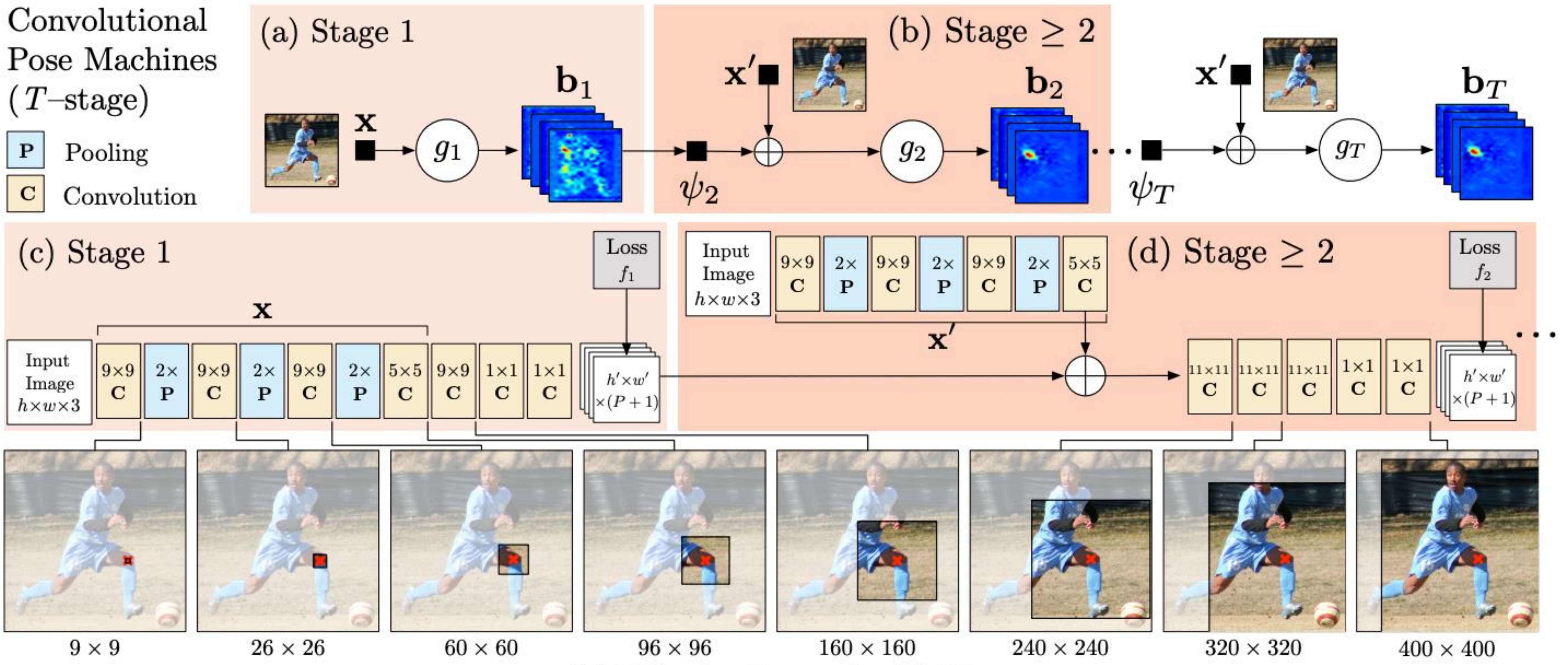


Target “belief map” :
K+1 x H x W
1 at Grount truth location (x,y) for k-th keypoint in the k-th channel

[Wei et al CVPR 2016]

Convolutional Pose Machines

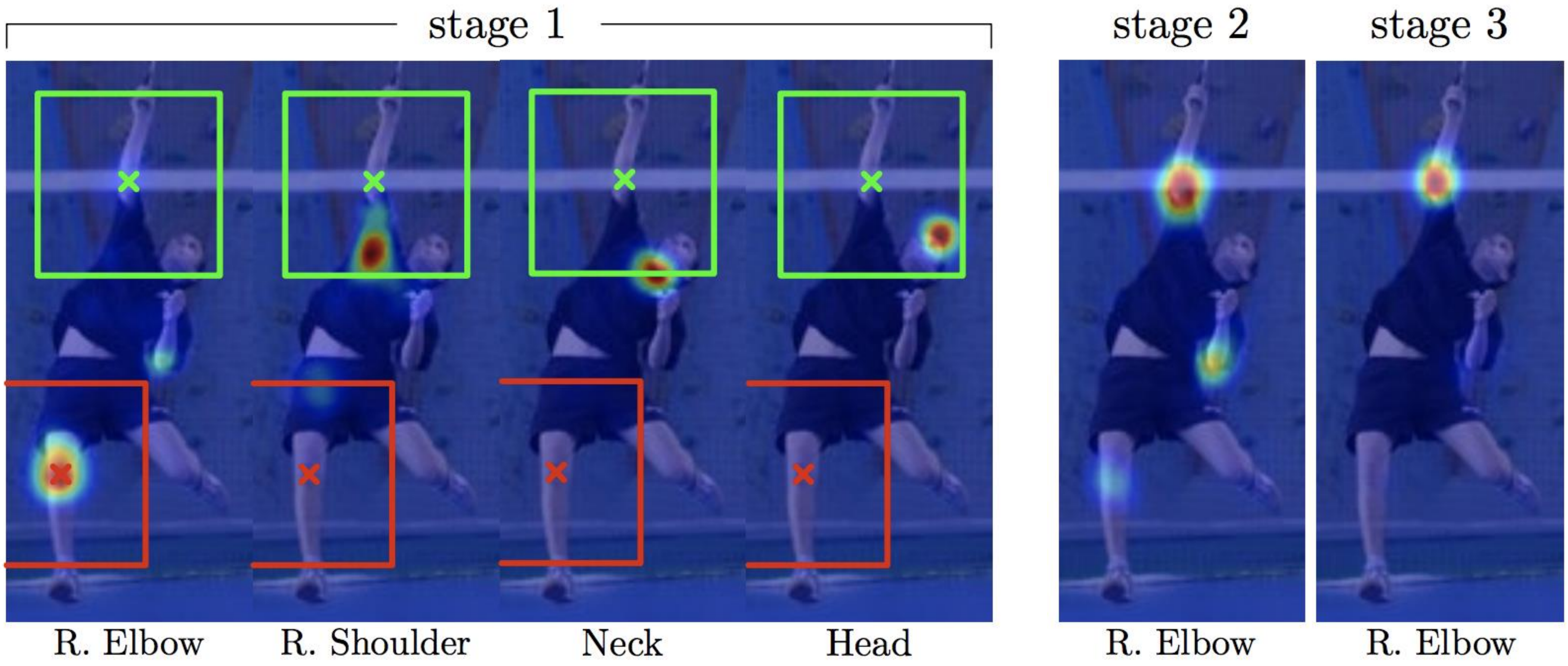
Base architecture for OpenPose



(e) Effective Receptive Field

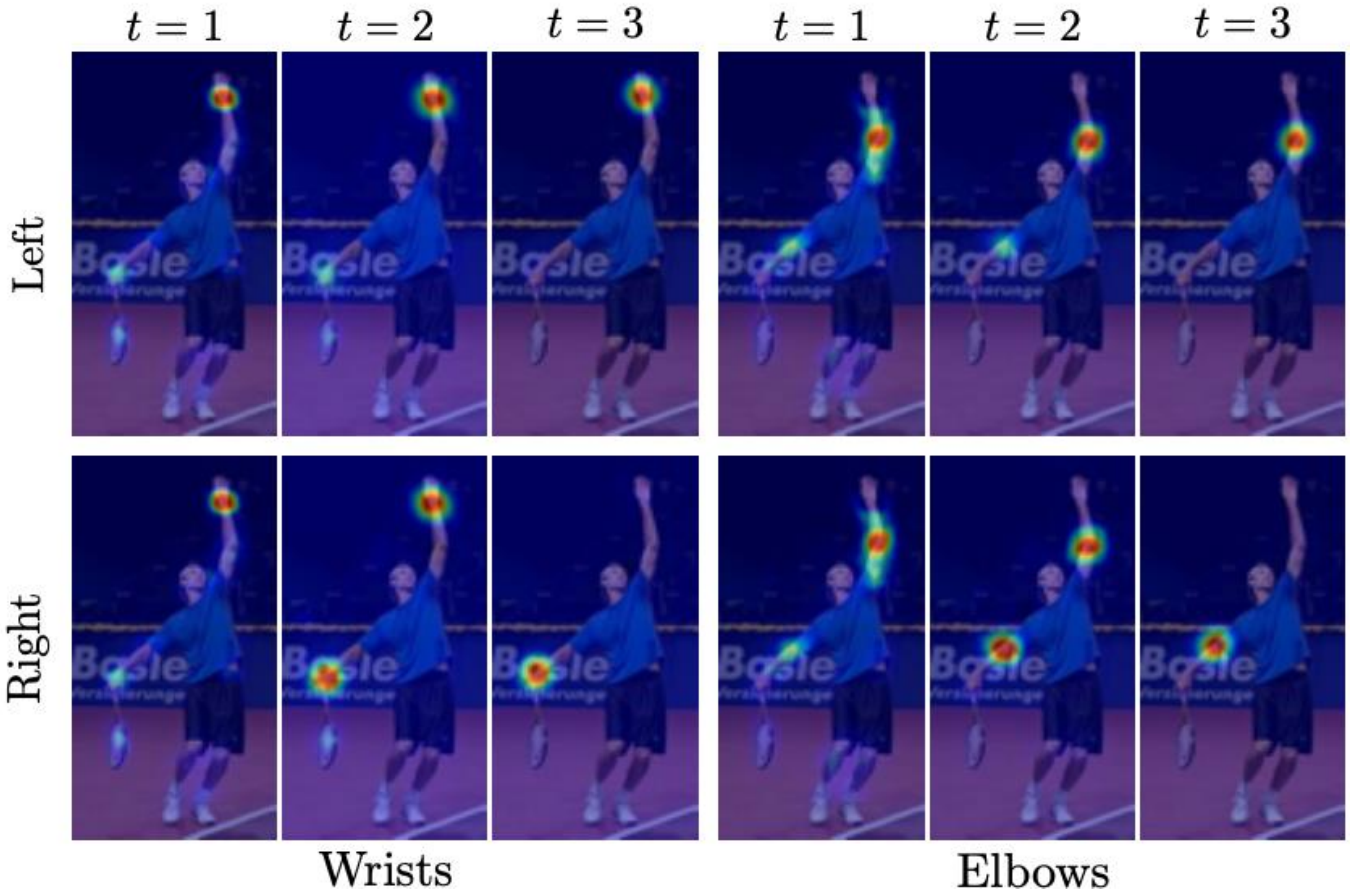
[Wei et al CVPR 201

Convolutional Pose Machines

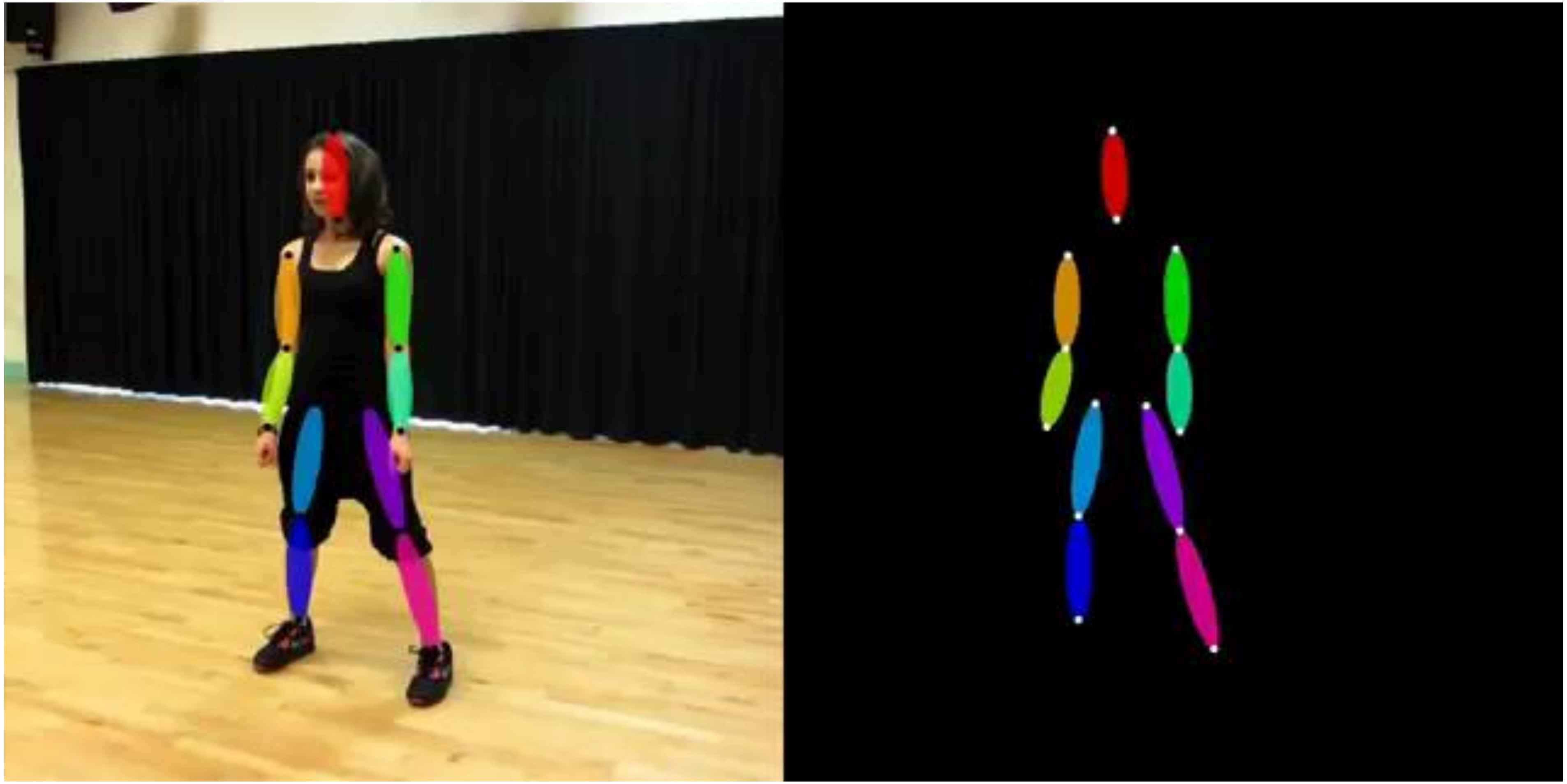


[Wei et al CVPR 2016]

Convolutional Pose Machines

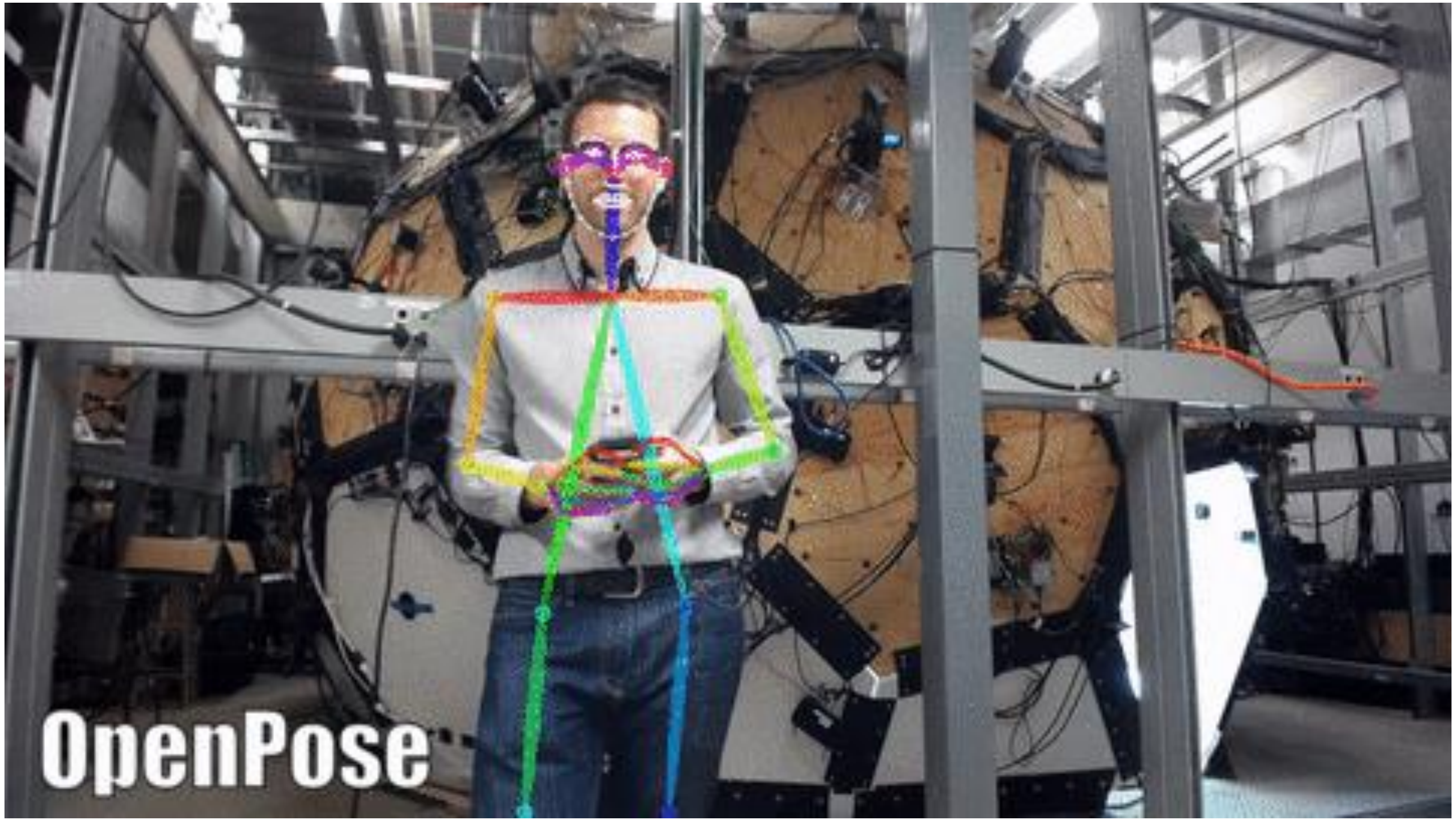


Results



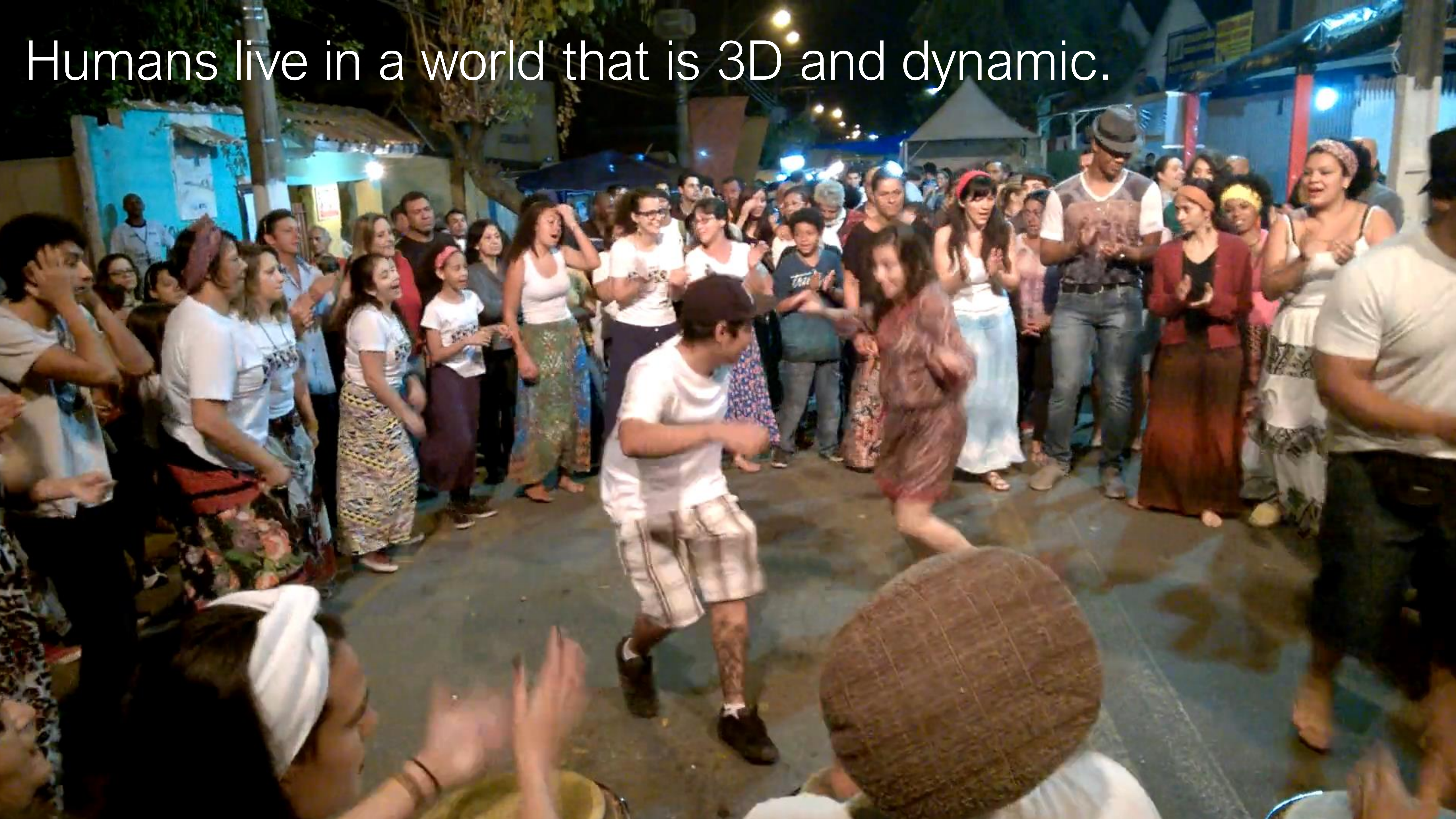
OpenPose

Great opensource tool, builds on convolutional pose machine architecture, adapted to multiple people



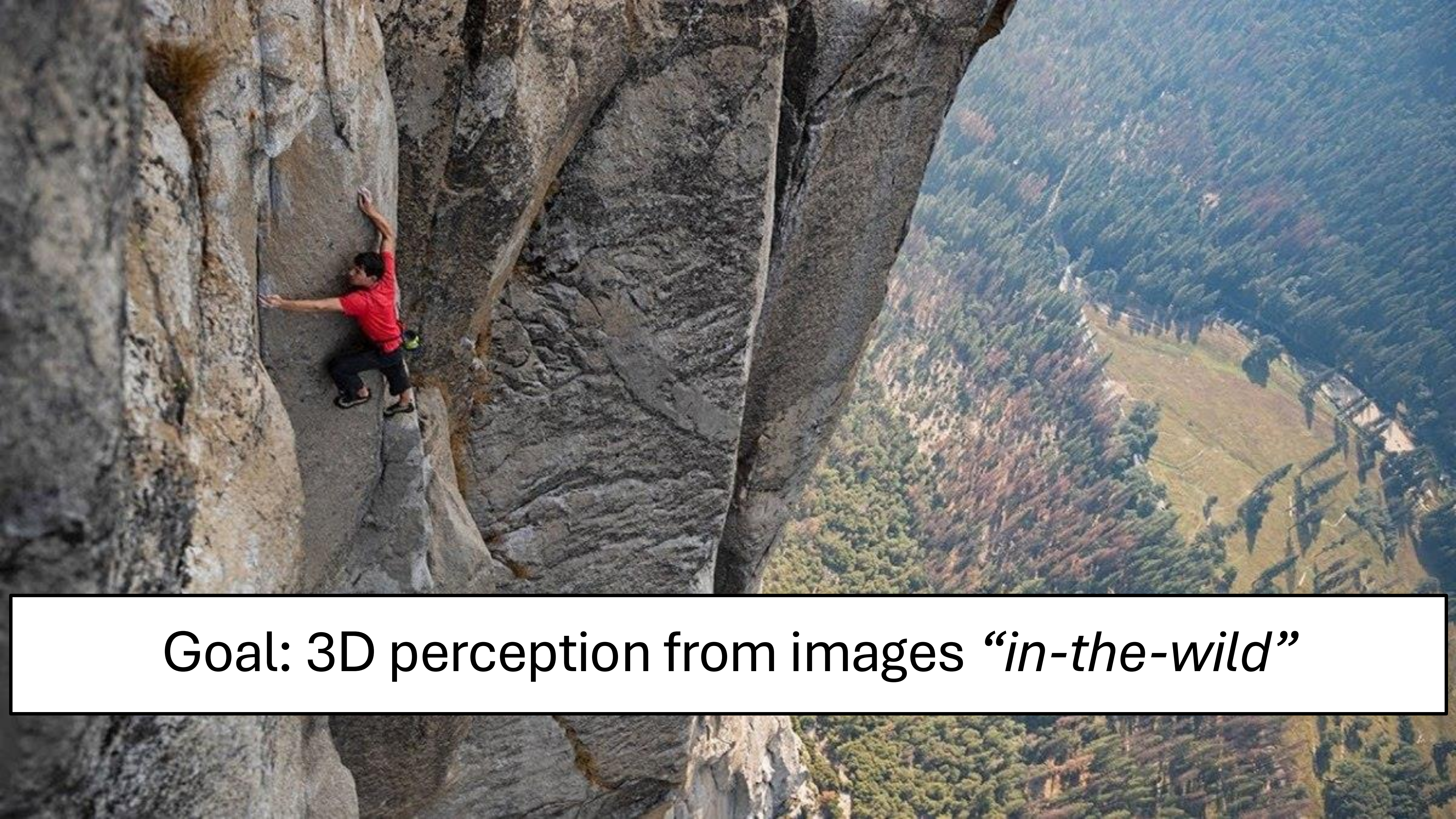
Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh ‘16-17

Are we done?


Humans live in a world that is 3D and dynamic.

Today’s Non-rigid 3D Solution: Motion Capture

The world is so much more than greenscreen!

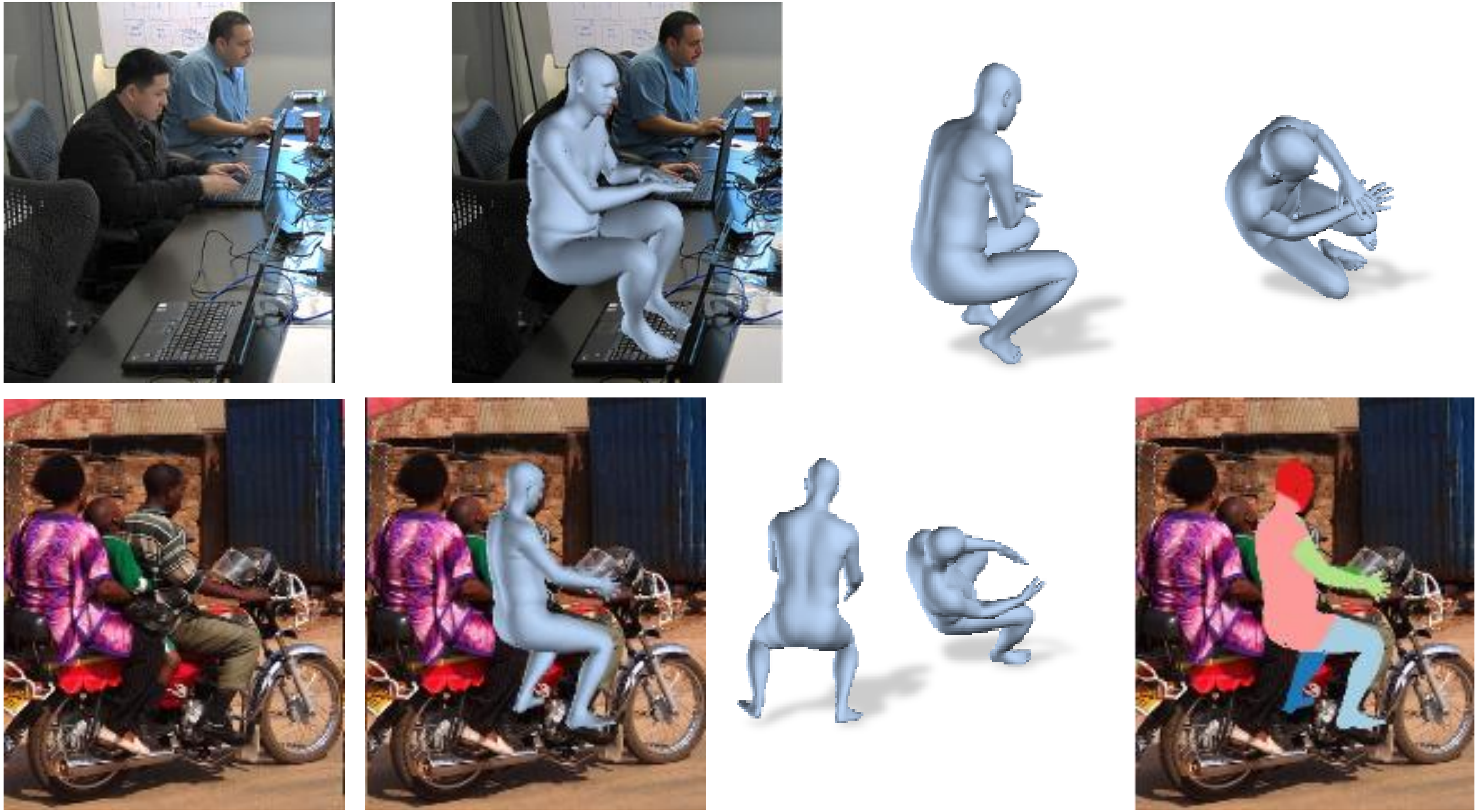


Goal: 3D perception from images *"in-the-wild"*

Single-View 3D Human Mesh Recovery

[Bogo∗, Kanazawa∗, Lassner, Gehler, Romero, Black ECCV ’16]

In everyday photos



Kanazawa, Black, Jacobs, Malik. Human Mesh Recovery, CVPR 2018

Or from Video

Kanazawa, Zhang, and Felsen et al. CVPR 2019

Learning to act from visual observation

From video…

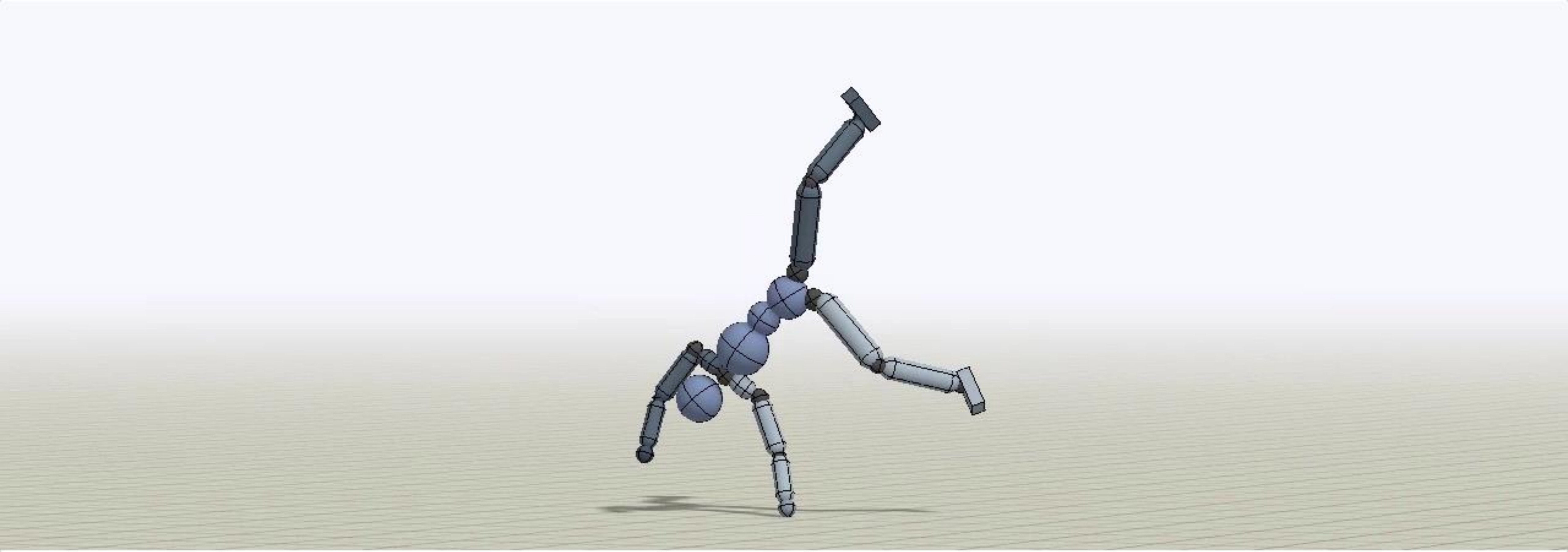
Video

Recovered 3D Body

Policy

Peng, Kanazawa, Malik, Abbeel, Levine *“SFV: Reinforcement Learning of Physical Skills from Videos”*, SIGGRAPH Asia 2018

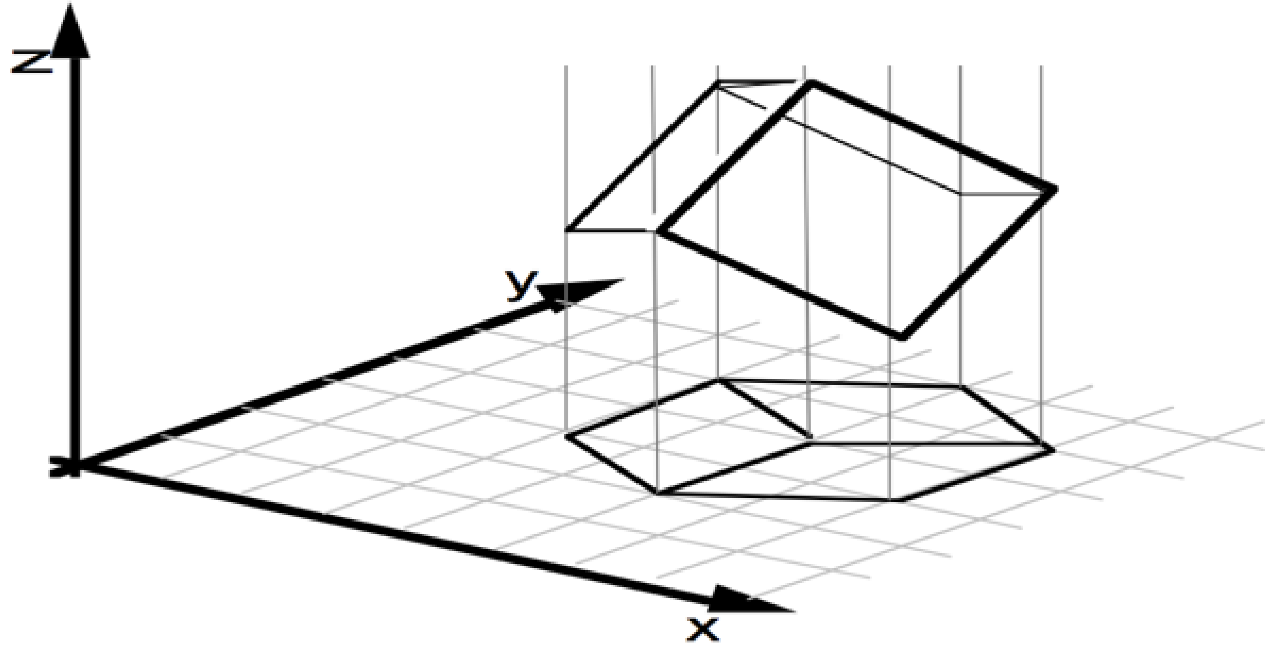
Animate Virtual Characters



Human 3D perception

Johannson experiment, James Maas, 1971

3D from 2D is inherently under-constrained



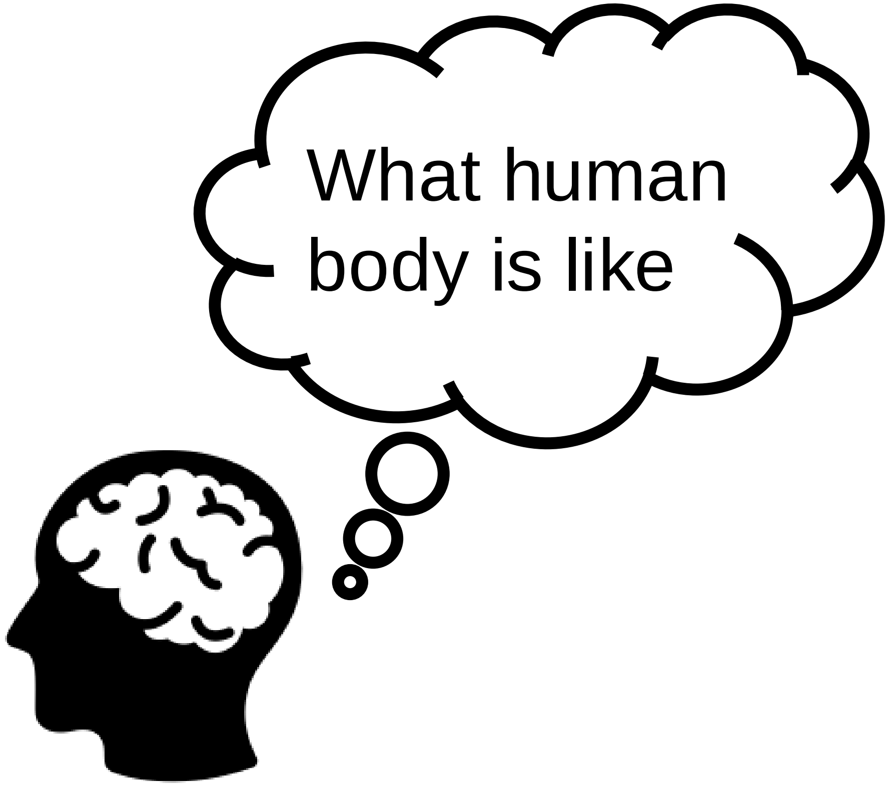
[Sinha and Adelson ‘93]

How do we resolve this?

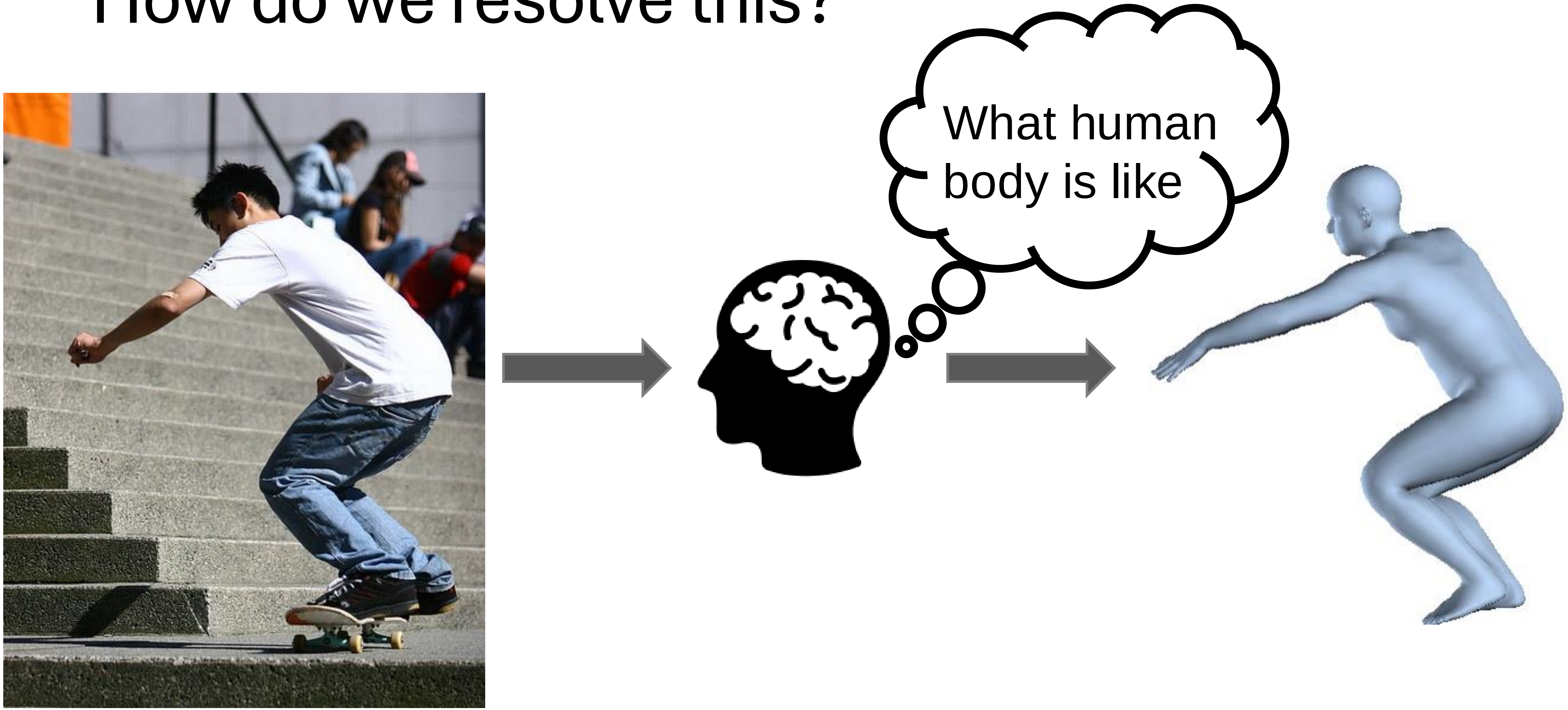
[Bogo and **Kanazawa** et al. ECCV ’16]

How do we resolve this?

“Remembrance of Objects Past”

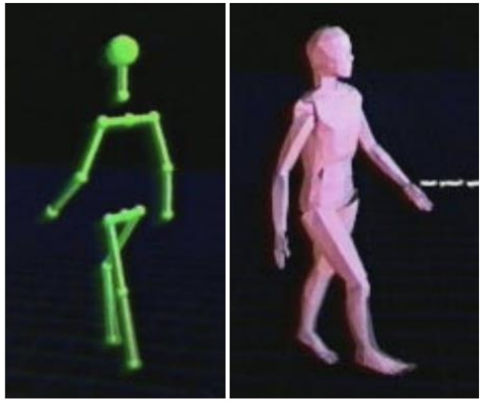


How do we resolve this?



[**Kanazawa** et al. CVPR 2018]

Bregler and Malik CVPR 1998

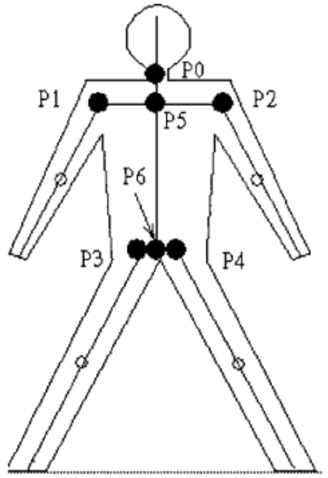


Tracking based:

1. Initialize 3D model in first frame
2. Track parts in next frames via Lukas-Kanade, over joint angles

More stable with 2 views

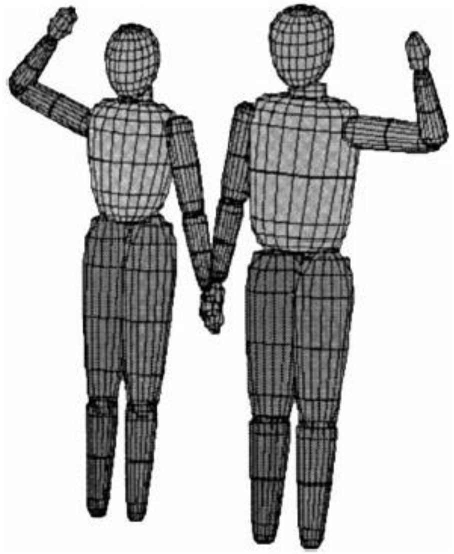
And many more model-based methods



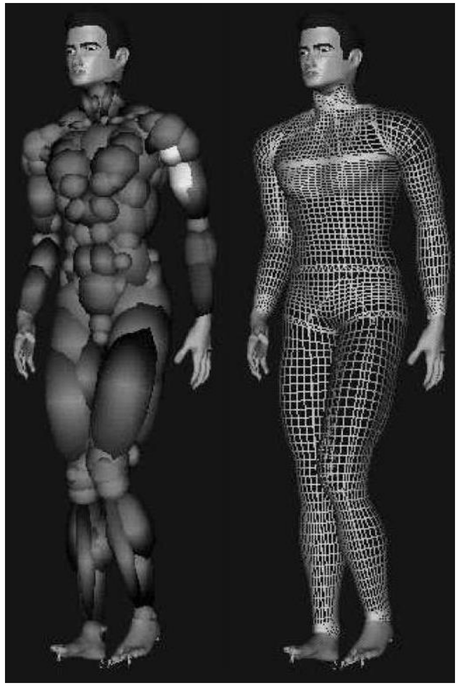
[Leung and Yang ‘95]

[Terzopoulos
and Metaxas ’93]

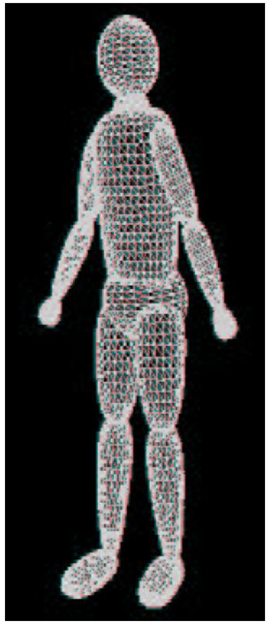
[Kakadiaris and Metaxas ’00]



[Gavrilla, ‘96]



[Plänkers and Fua ’01
]



[
Sminchisescu
and Triggs ’03]

Slide modified from Michael
Black

3D Humans from known 2D joints

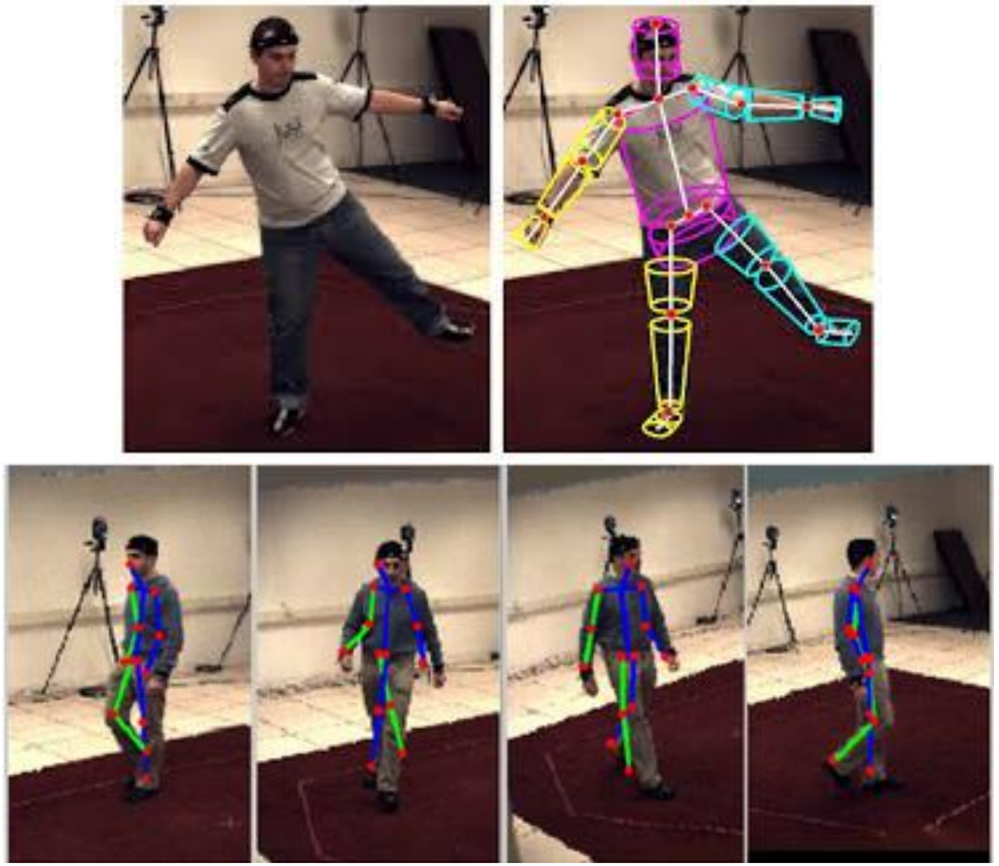
Reconstruction of articulated objects from point correspondences in single uncalibrated image
[CJ Taylor CVIU 2000]

Same issue as single-view 3D reconstruction

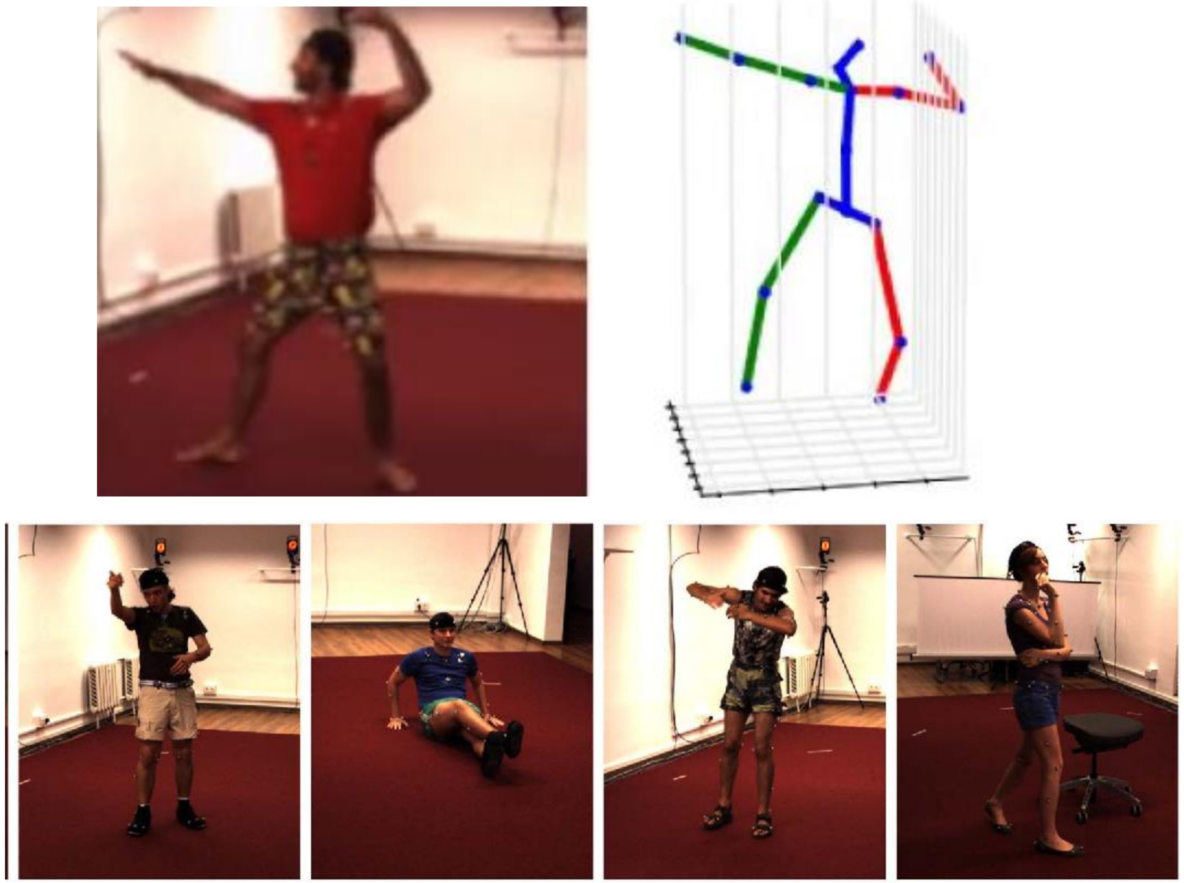
(a)

You need priors!!
Here: Known ratio of limb length

Datasets



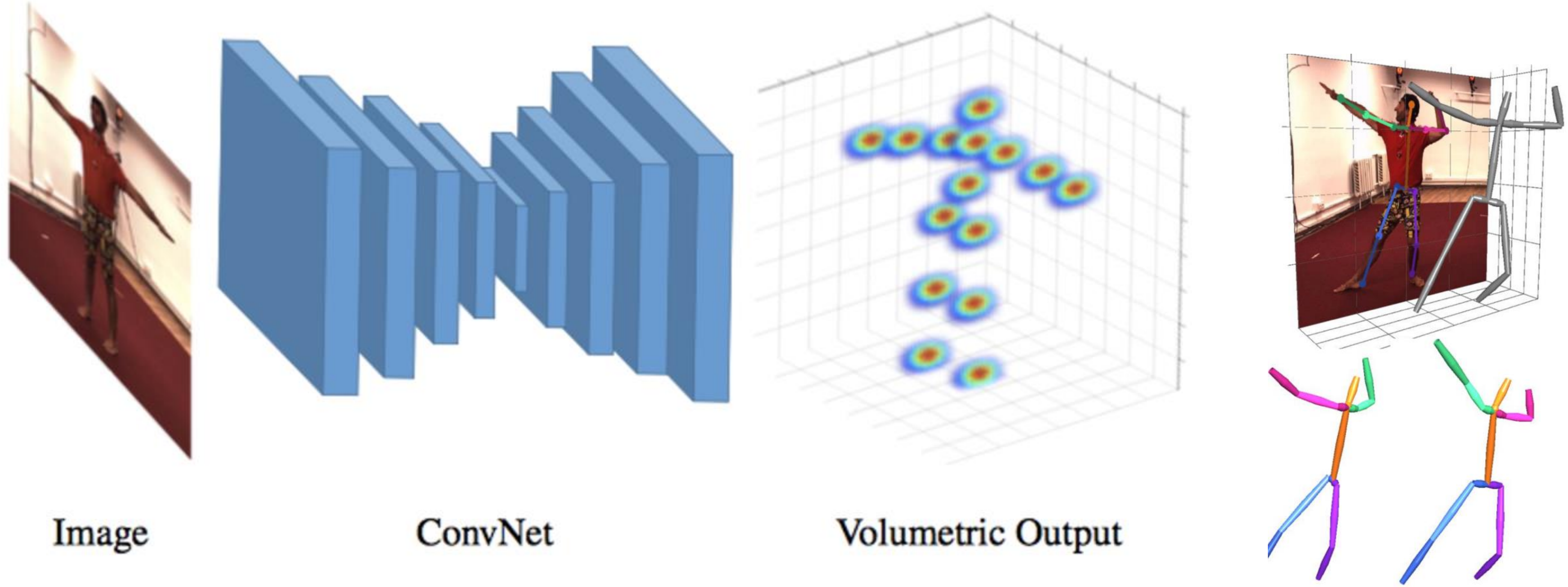
HumanEva [Sigal et al. IJCV 2010]



Human3.6M [Ionescu et al. 2014]

Deep Learning based approaches

Lots of activities + progress made in this area after datasets

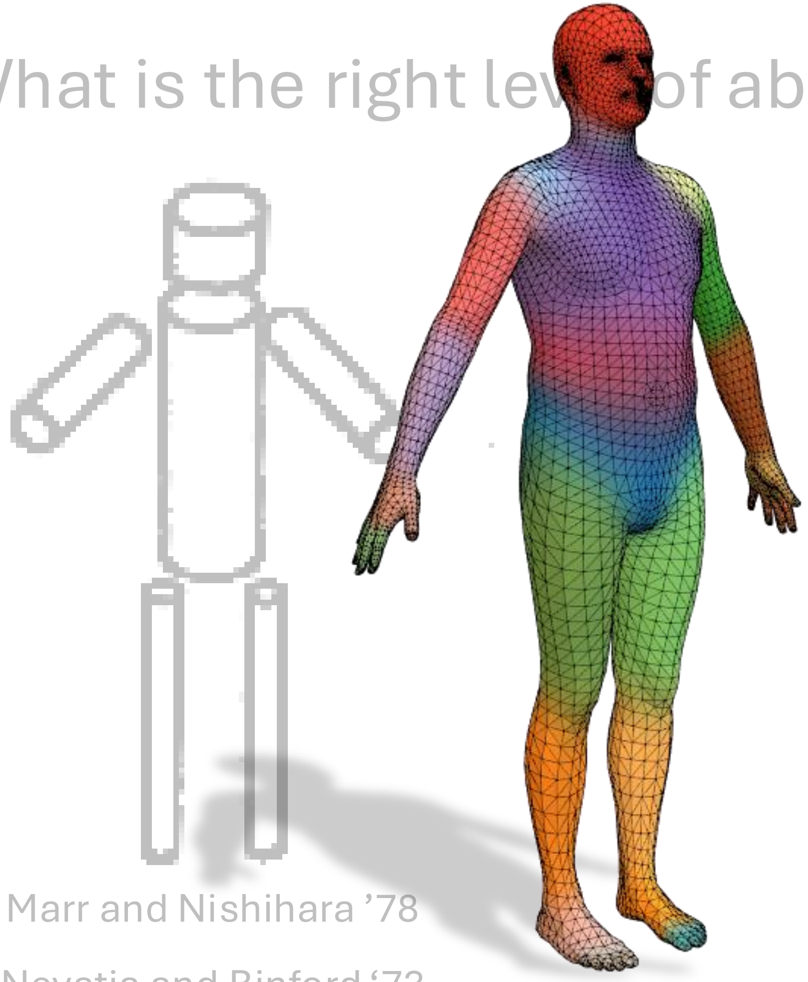


[Pavlakos et al. CVPR'17, Sun et al ICCV'17, Vnect Mehta et al SIGGRAPH '17 ...]

What about the 3D representation??
Are we all just stick figures?

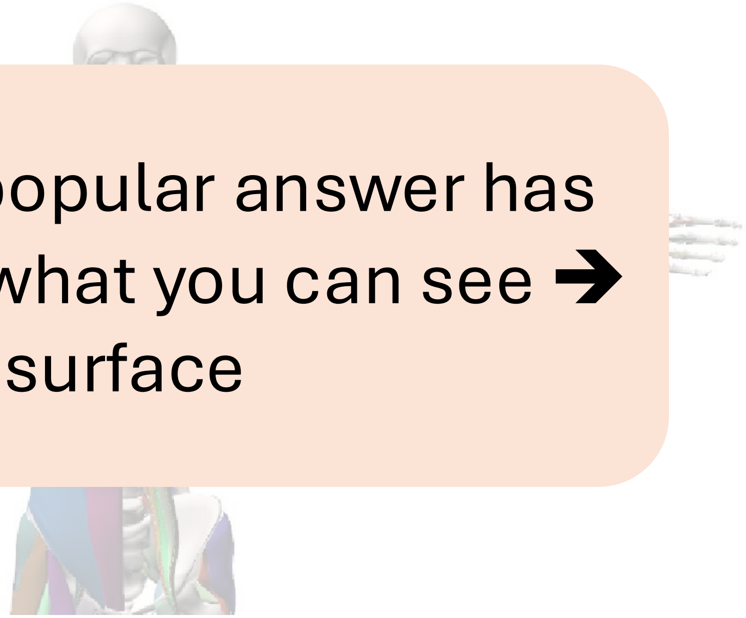
To do more than joints, we need to discuss how to model human bodies

What is the right lev of abstraction?



Marr and Nishihara '78

Nevatia and Binford '73

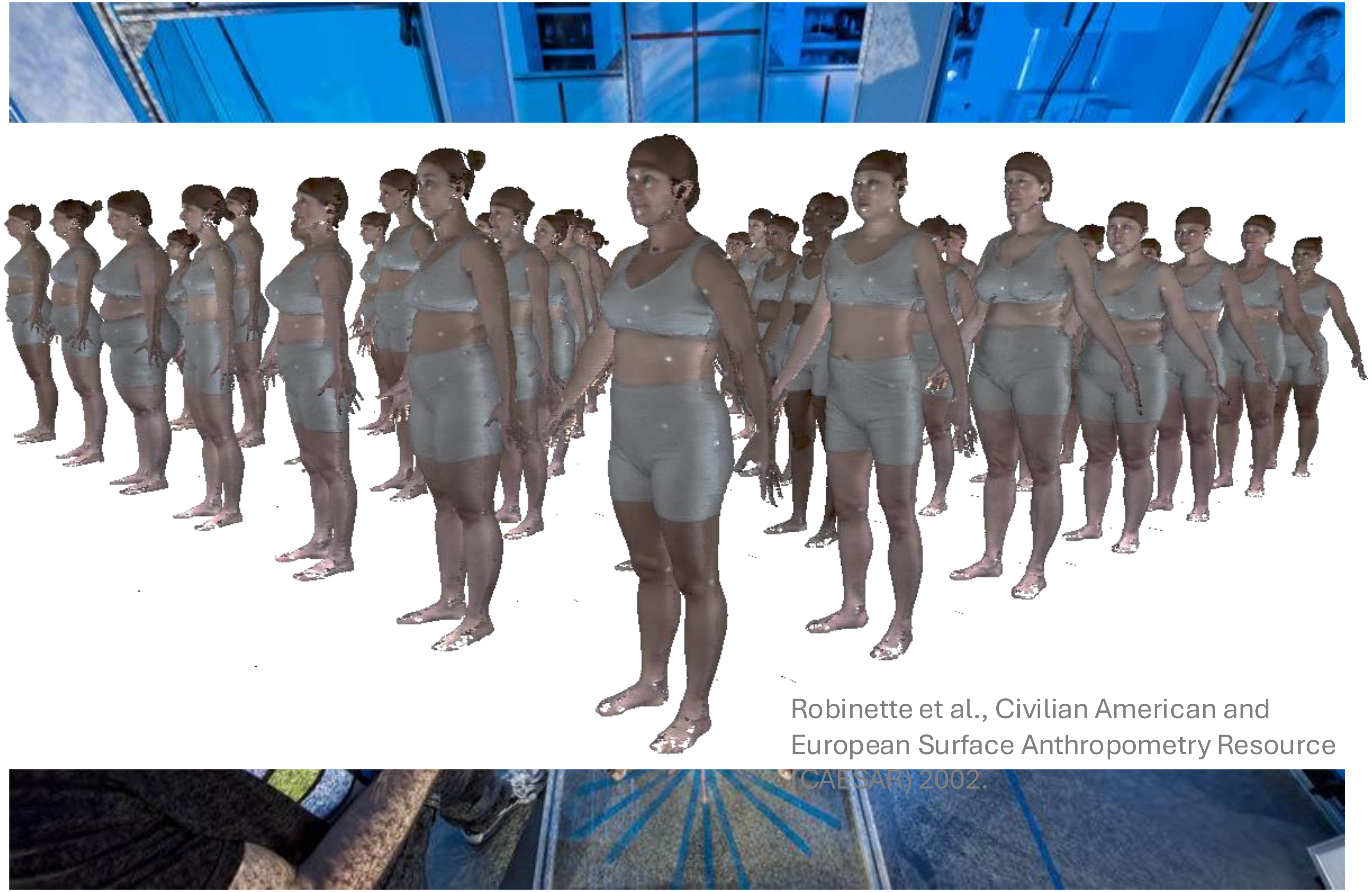


Lee, Sifakis, Terzopoulos, ToG'09

Practical and popular answer has been to model what you can see ➔ the surface

Slide courtesy of Michael Black

Humans are special



Robinette et al., Civilian American and European Surface Anthropometry Resource (CAESAR) 2002.

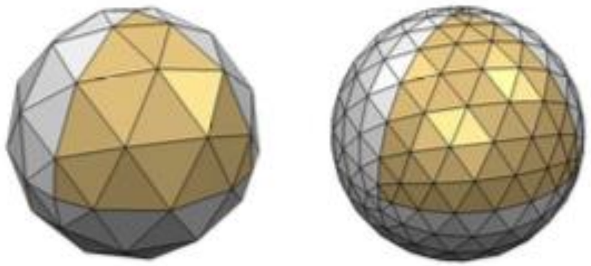
Morphable Model of Human Bodies

[Loper et al. SIGGRAPH Asia ‘15]

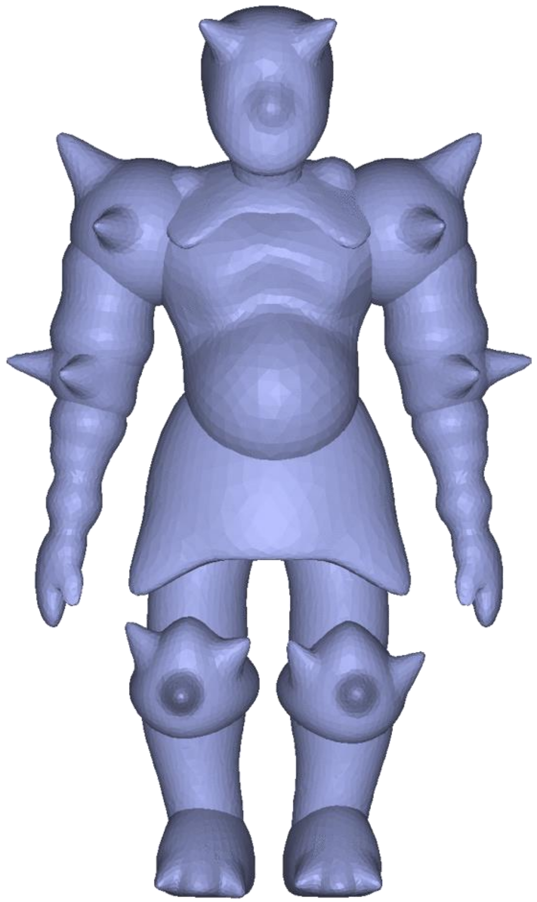
How to represent surfaces?

We need a low-dimensional parametrization!!

- Meshes are a popular, practical choice for surfaces
- Mesh = {V, F}
  - Vertices: N x 3
  - Faces: |F| x {3, 4, …} polygons, "triangles"

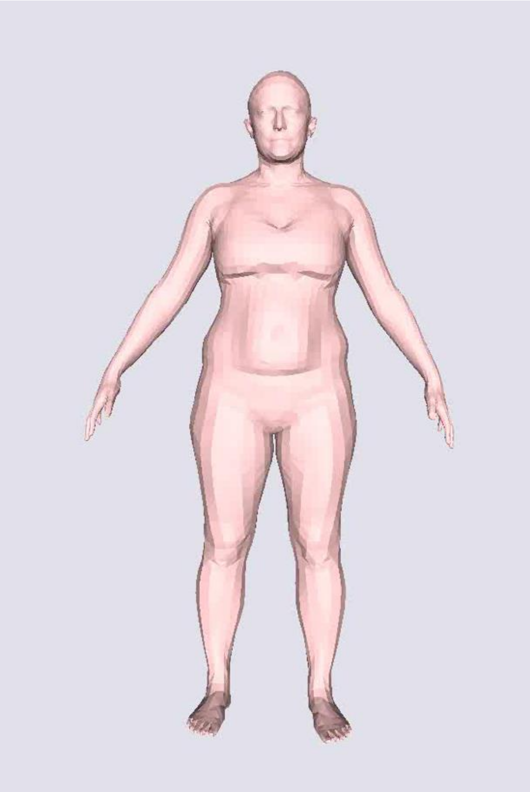


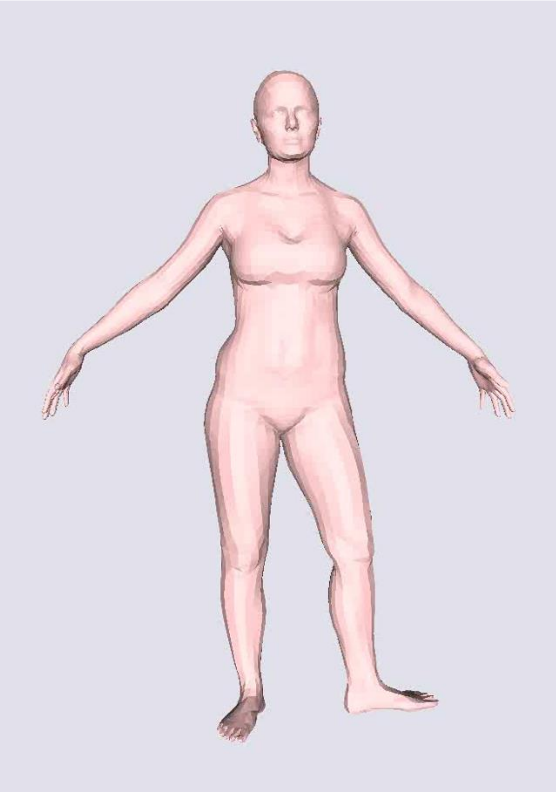
How to parametrize a mesh?



Key in modelling 3D Human Surfaces: Factorization into Shape and Pose

“Identity”





Individual Shape Variation

Pose changes (Articulation)

[SCAPE: Anguelov et al., SIGGRAPH ‘05]

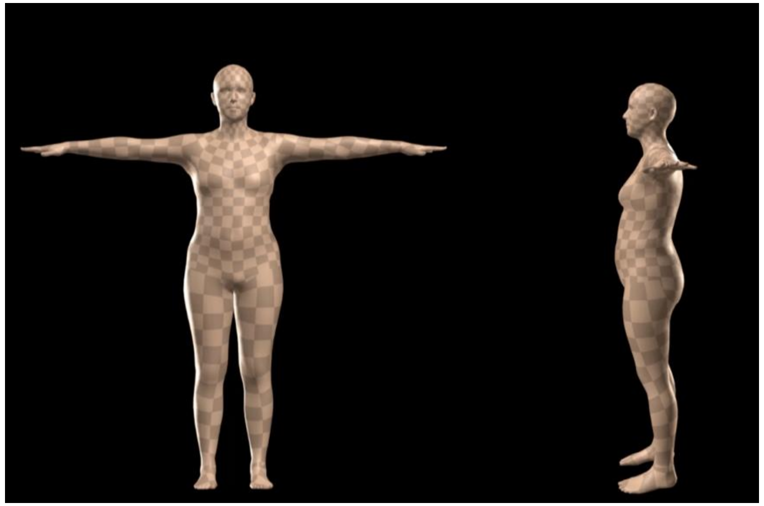
Figures courtesy of Michael Black

Skinned Multi-Person Linear Model (SMPL)

Shape: PCA coefficients

Pose: Rotation of joints

Mesh



$+$

$=$

$\vec{\beta}$ 10 dimensions

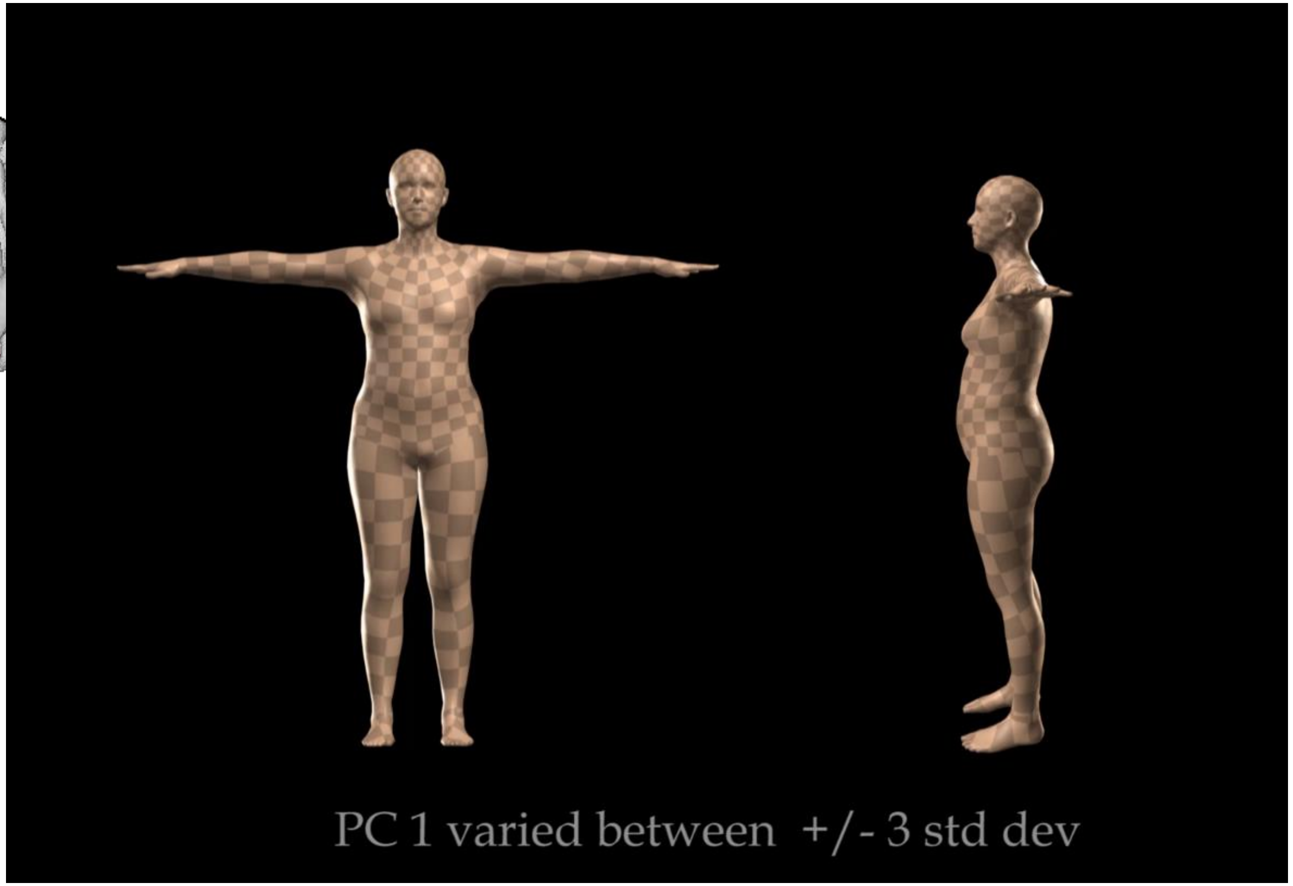
$\vec{\theta}$ 3*23 = 69 dimensions

SMPL [Loper et al. SIGGRAPH ASIA '16]

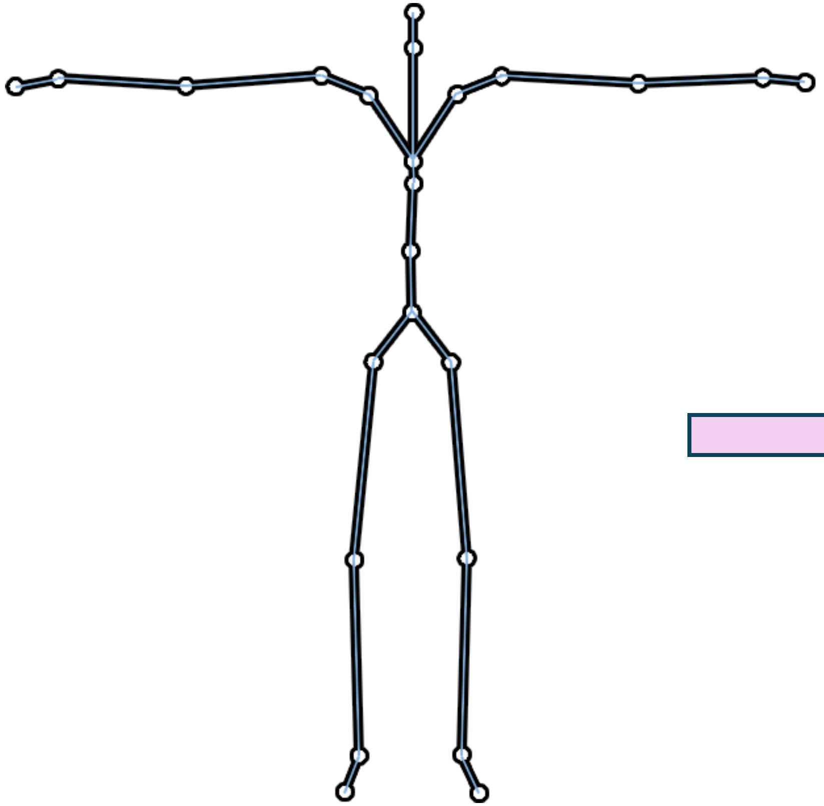
Learning Shape from 3D Scans

4000 bodies of different shapes in roughly the same pose.

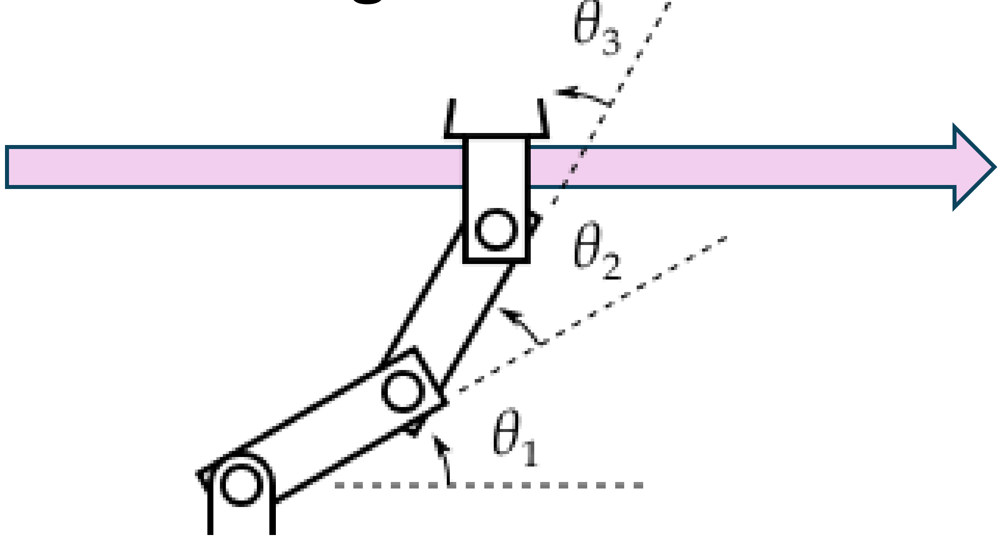
Run PCA on this:
Shape = linear combination of basis shapes

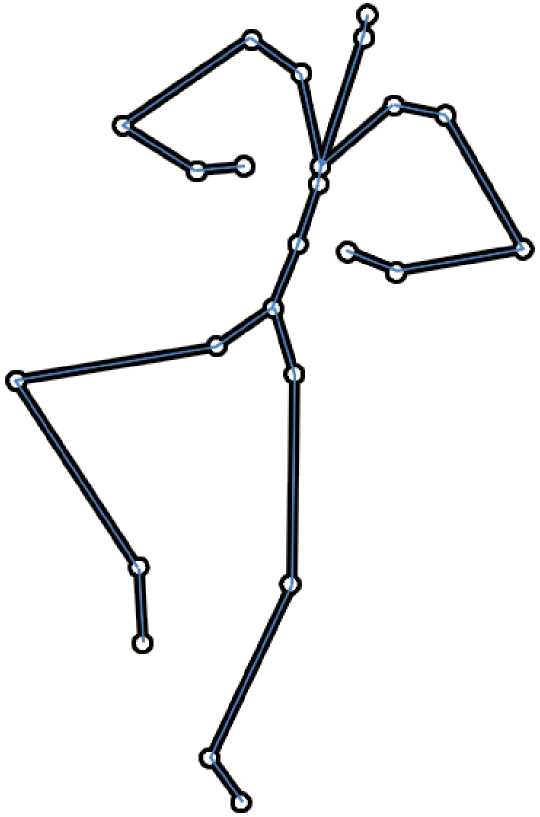


Pose: Forward kinematics on the skeleton tree



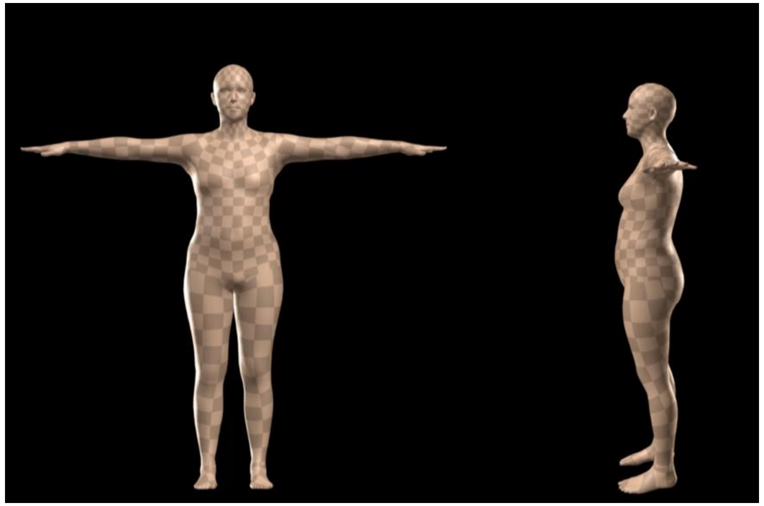
Defined by relative joint angles.





Skinned Multi-Person Linear Model (SMPL)

Shape: PCA coefficients



$\vec{\beta}$ 10 dimensions

$+$

Pose: Rotation of joints

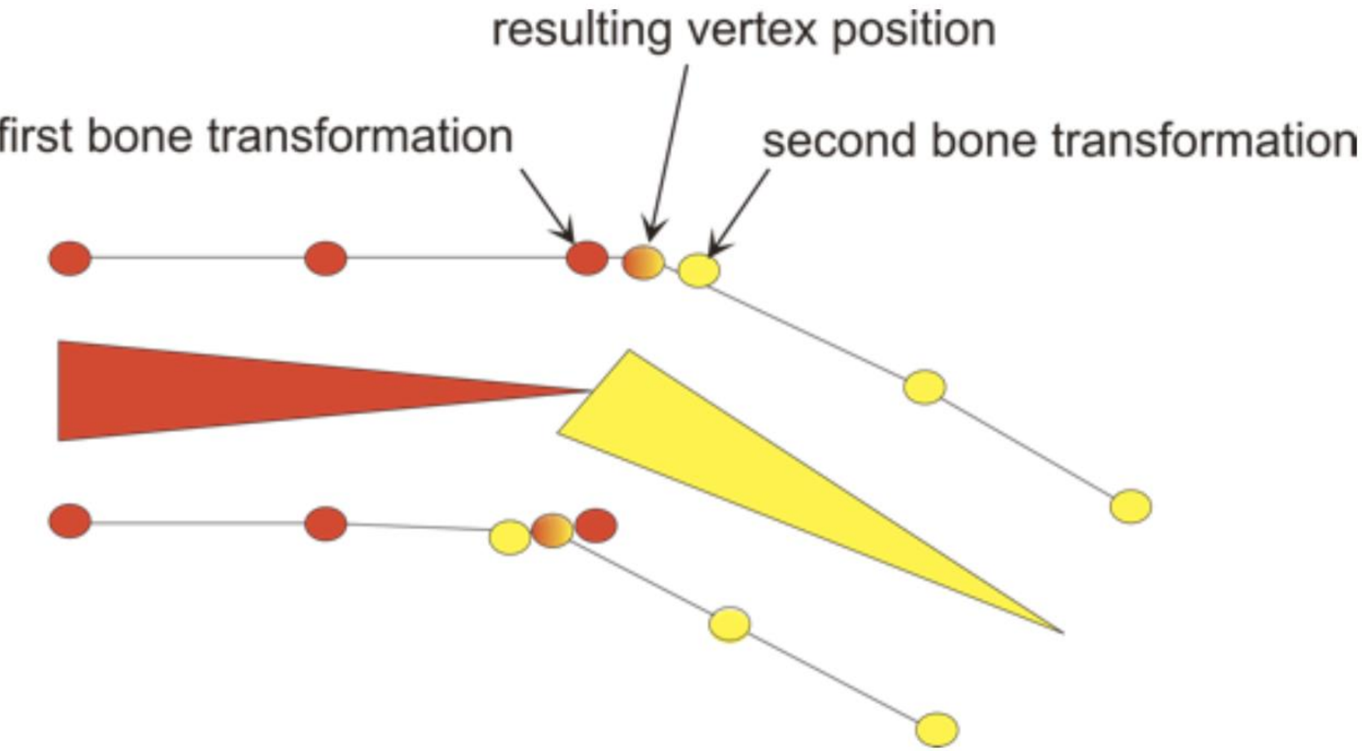
$\vec{\theta}$ 3*23 = 69 dimensions

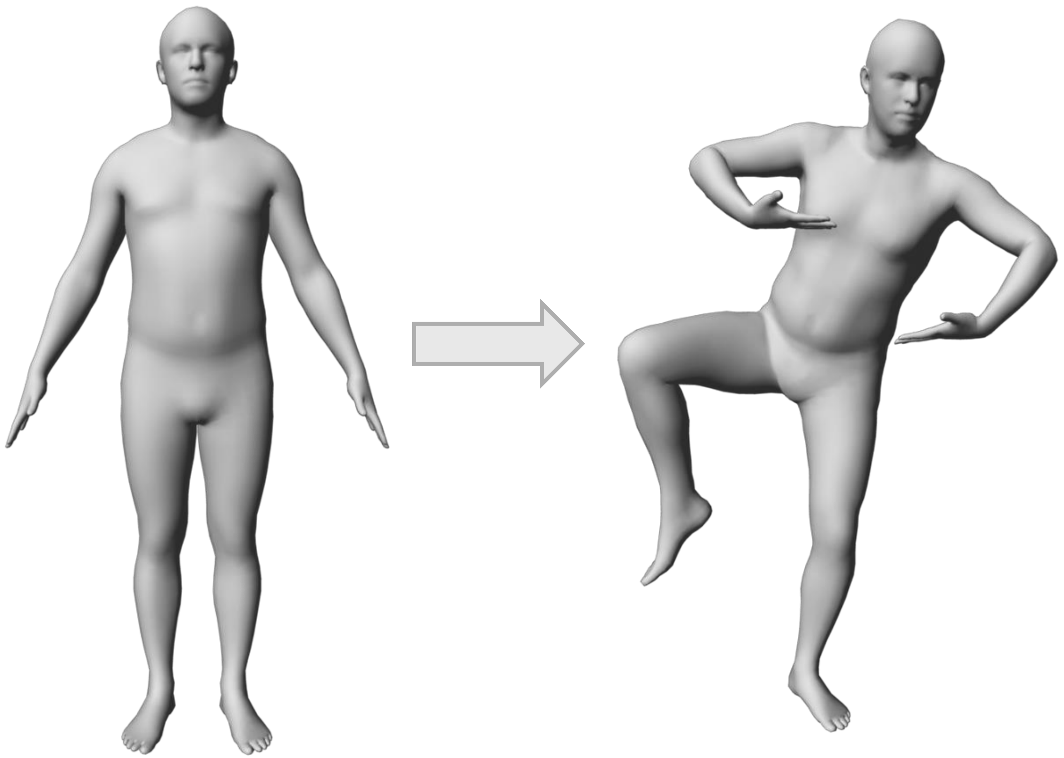
$=$

Mesh

SMPL [Loper et al. SIGGRAPH ASIA '16]

Pose: Forward kinematics on the skeleton tree



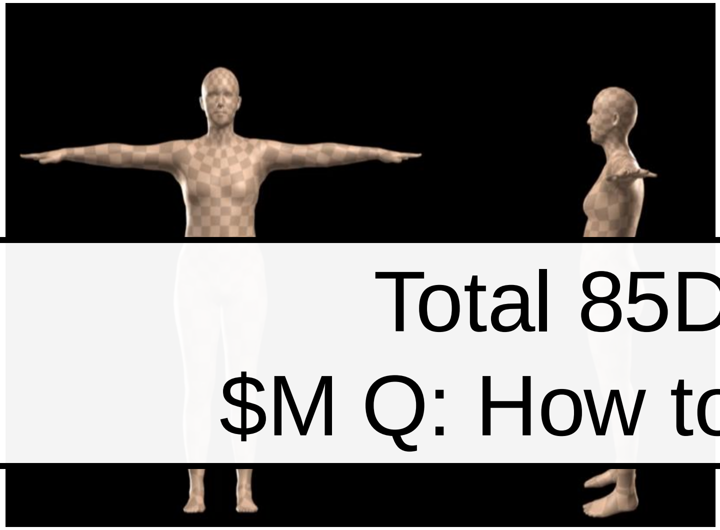


Morphable Model of Human Bodies

[Loper et al. SIGGRAPH Asia '15]

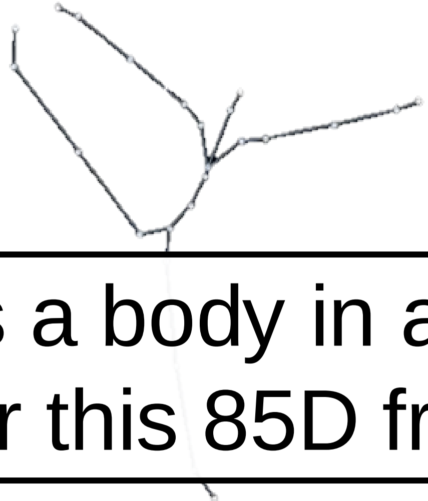
Morphable model for humans

Shape: low-D subspace



$+$

Pose: 23 Joint Rotations



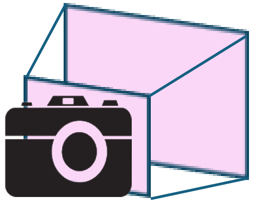
$=$

Mesh

Total 85D defines a body in an image!!!
$M Q: How to recover this 85D from an image?

$\vec{\beta} \in \mathbb{R}^{10}$

$\vec{\theta} \in \mathbb{R}^{23 \times 3}$



$\Pi \in \mathbb{R}^{6}$

Skinned Multi Person Linear (SMPL) model [Loper et al. SIGGRAPH ASIA ‘15]

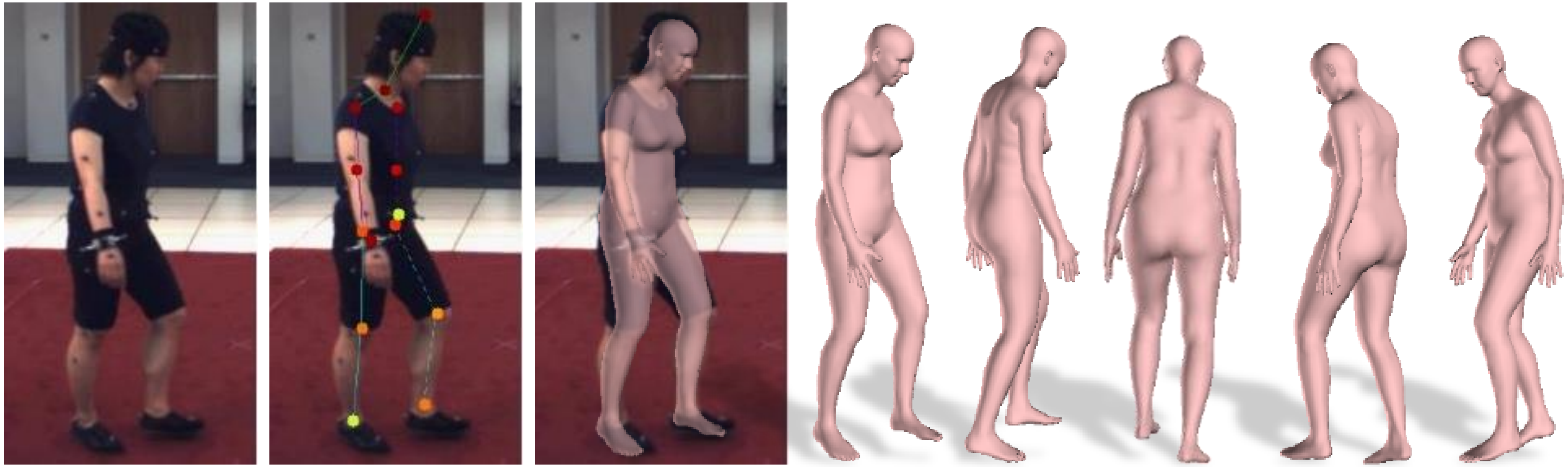
Back to images…
3D Shape and Pose from a Single Image

SMPLify [Bogo and Kanazawa et al ECCV'12]

Overview: SMPLify

1. Automatic 2D joint detection via CNNs

2. Fit SMPL pose and shape parameters



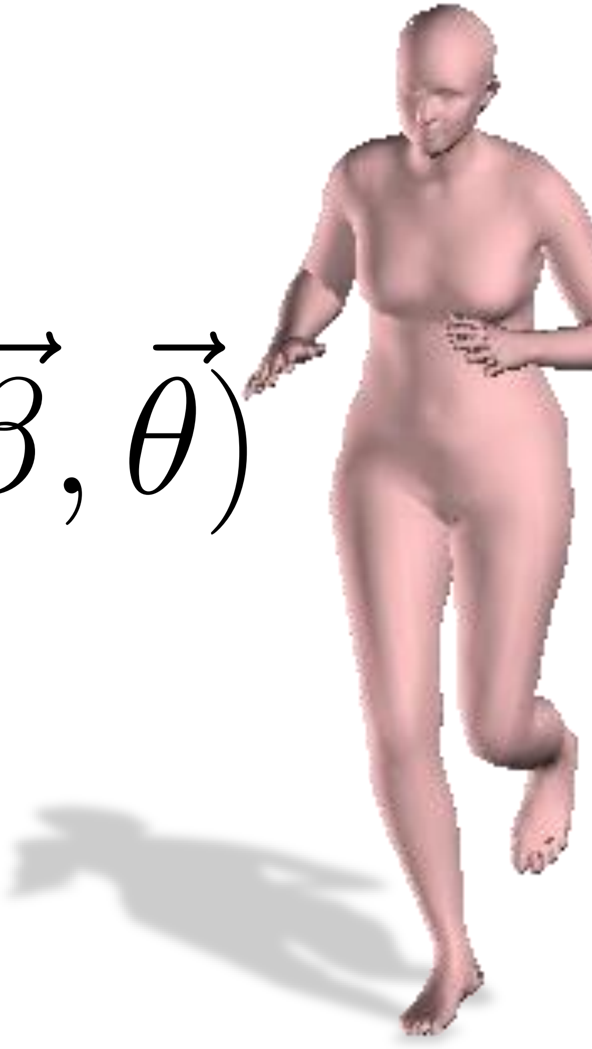
SMPLify Objective Function

camera joints

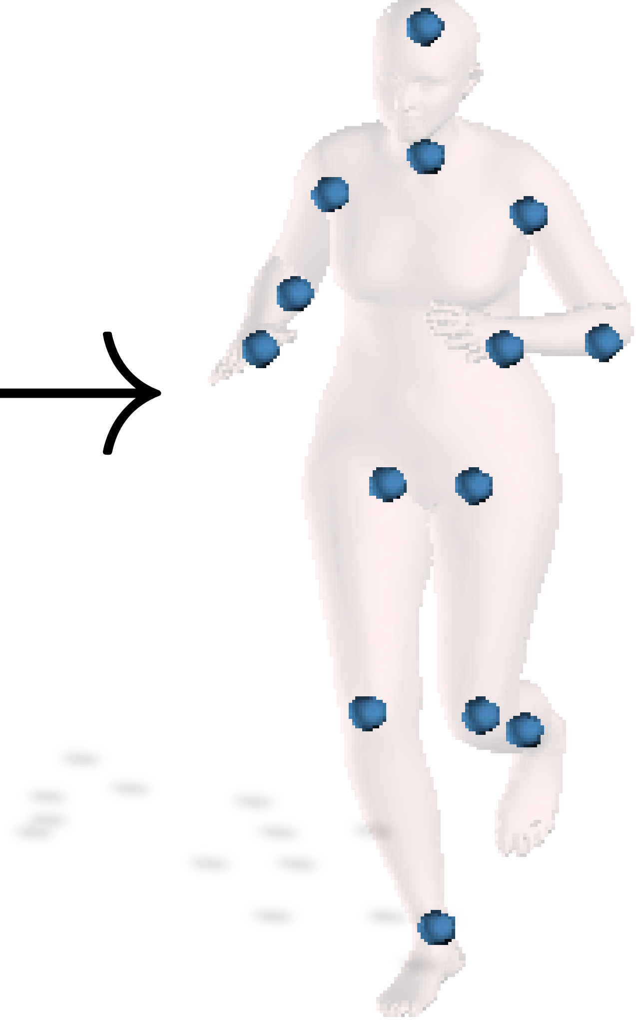
$$E(\vec{\beta}, \vec{\theta}, K; J_{est}) =$$

$$\underbrace{E_J(\vec{\beta}, \vec{\theta}, K; J_{est})}_{\text{Data term}} + \underbrace{E_a(\vec{\theta}) + E_\theta(\vec{\theta}) + E_{sp}(\vec{\theta}, \vec{\beta}) + E_\beta(\vec{\beta})}_{\text{Priors}}$$

Data Term: Joint Reprojection Error

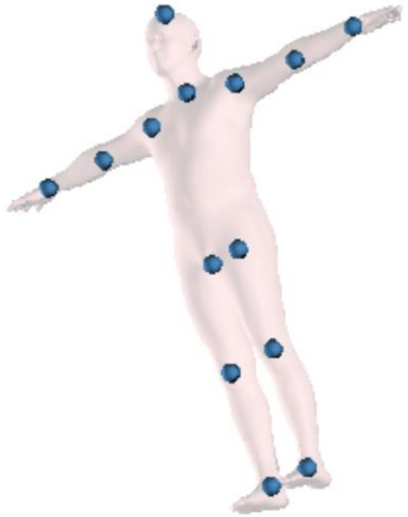


$$(\vec{\beta}, \vec{\theta}) \longmapsto$$



Data Term: Joint Reprojection Error

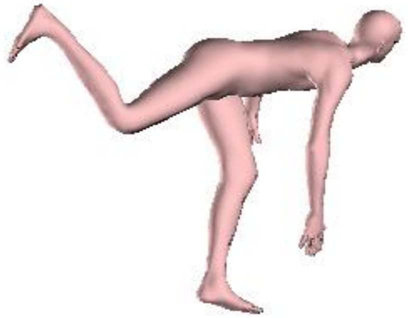
Camera Projection

$\Big\|$  $-\Pi_K\Big($  $\Big)\Big\|_2^2$

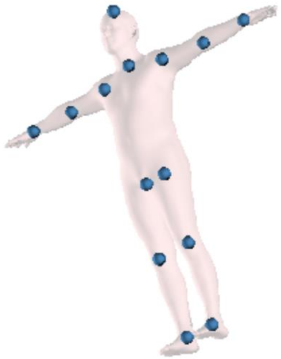
Summary: Fit to 2D joints

1. Automatic 2D joint detection via CNN

2. Solve for pose and shape that explain the 2D joints

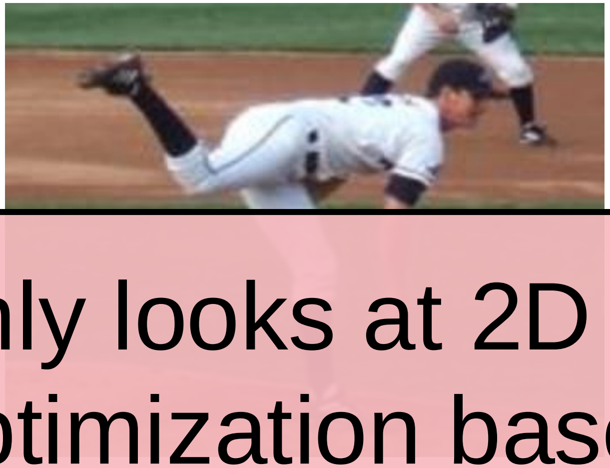


$\min_{\beta,\theta,\Pi} \Big\|$ 

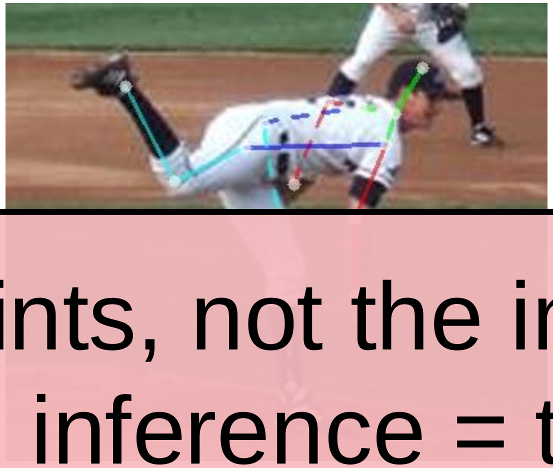
$- \ \Pi($  $) \Big\|_2^2 \ +$ lots of priors

[Bogo and **Kanazawa** et al ECCV ’16]

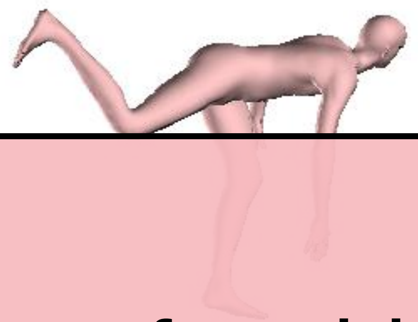
Approach: Fit to 2D joints



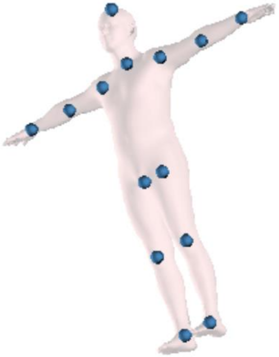
1. Automatic 2D joint detection via CNN



2. Solve for pose and shape that explain the 2D joints



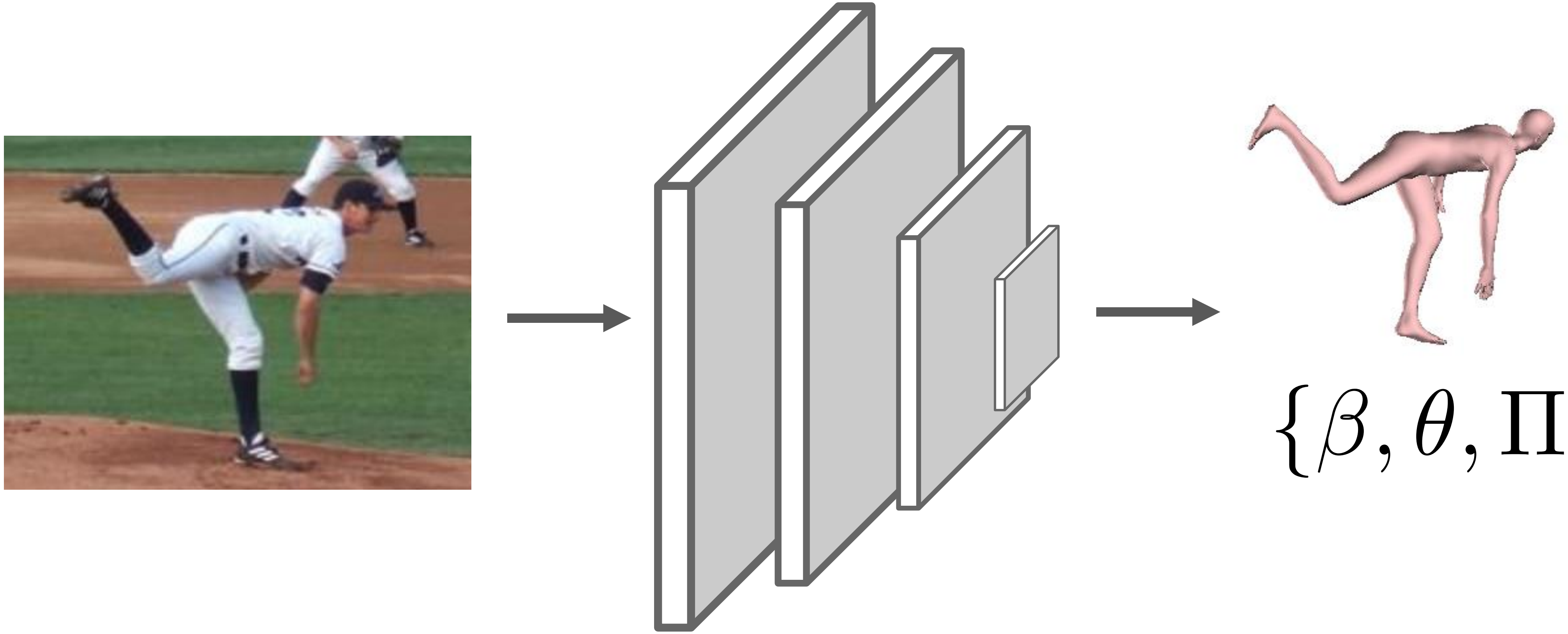
Only looks at 2D joints, not the image
Optimization based inference = too slow for video

$\min_{\beta,\theta,\Pi} \Big\|$  $- \Pi($  $) \Big\|_2^2 +$ lots of priors

[Bogo and **Kanazawa** et al ECCV ’16]

Why not just throw a deep network at it?

- Image in, 85D human parameters out!!



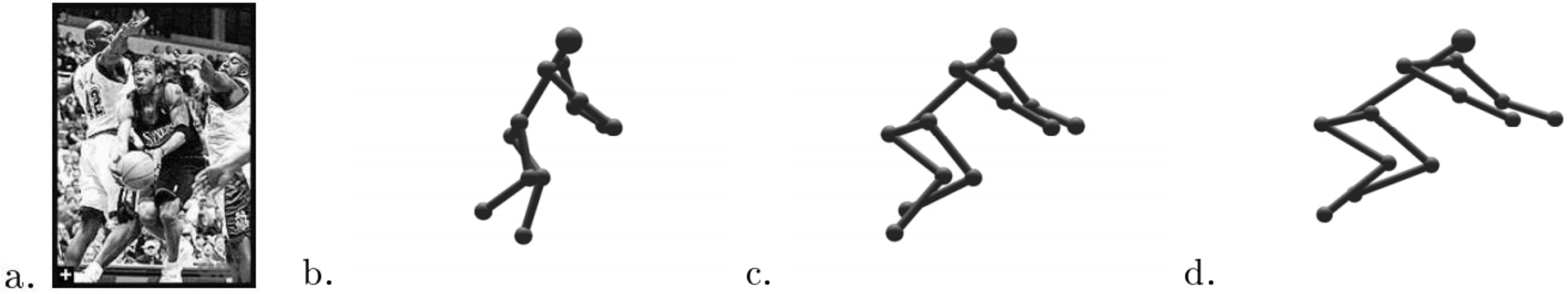
$$\{\beta, \theta, \Pi\}$$

Challenges

1. Lack of real paired 2D-to-3D labels

Human3.6M
[Ionescu et al.
PAMI ’14]

2. Depth ambiguity



[CJ Taylor CVPR 2000]

Solution

Even though we don't have paired 2D-to-3D labels,

we have a lot of **unpaired** labels

Explain the 2D

2D Labeled images

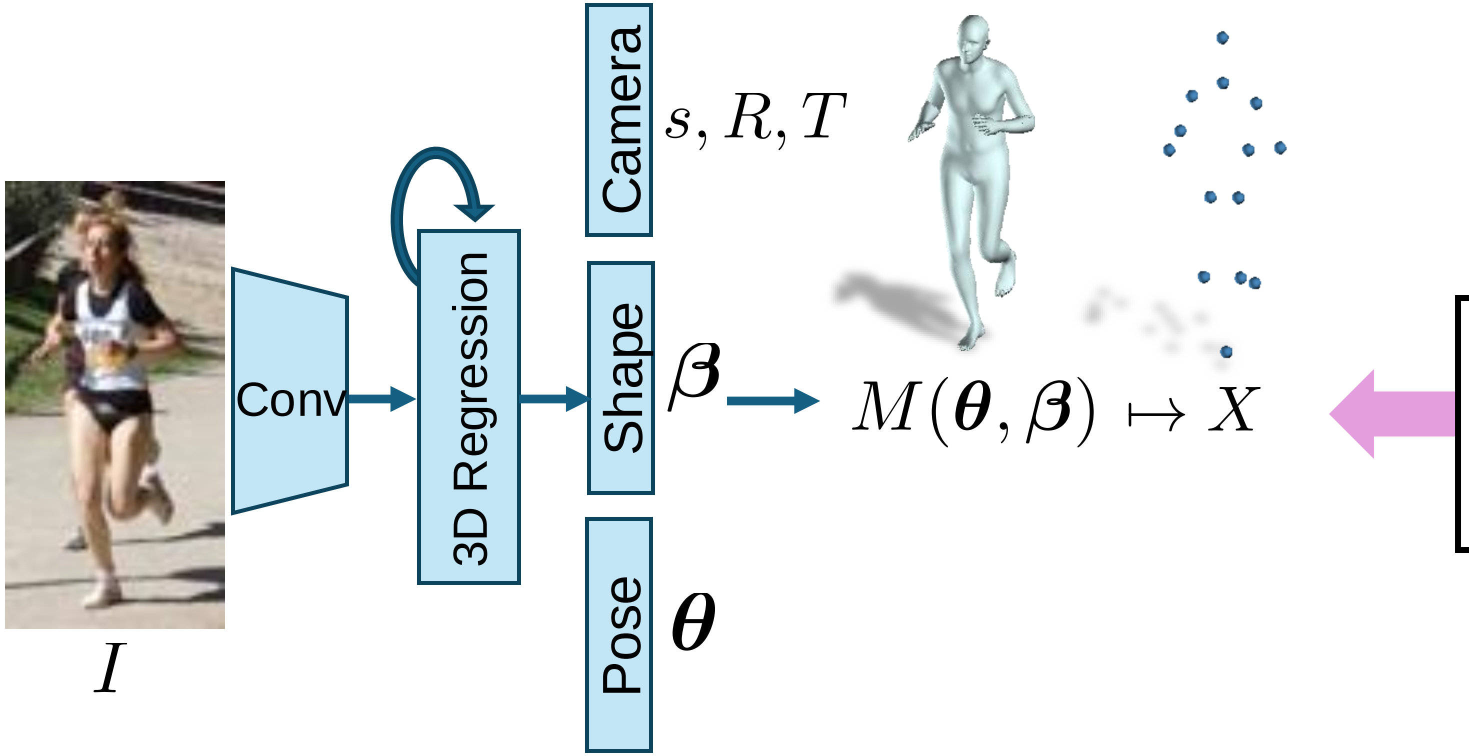
[LSP, MPII, MS COCO,…]

Within this distribution

3D Scans/Motion Capture

[CMU Mocap, CAESER, JointLimits..]

Overview: Human Mesh Recovery (HMR)

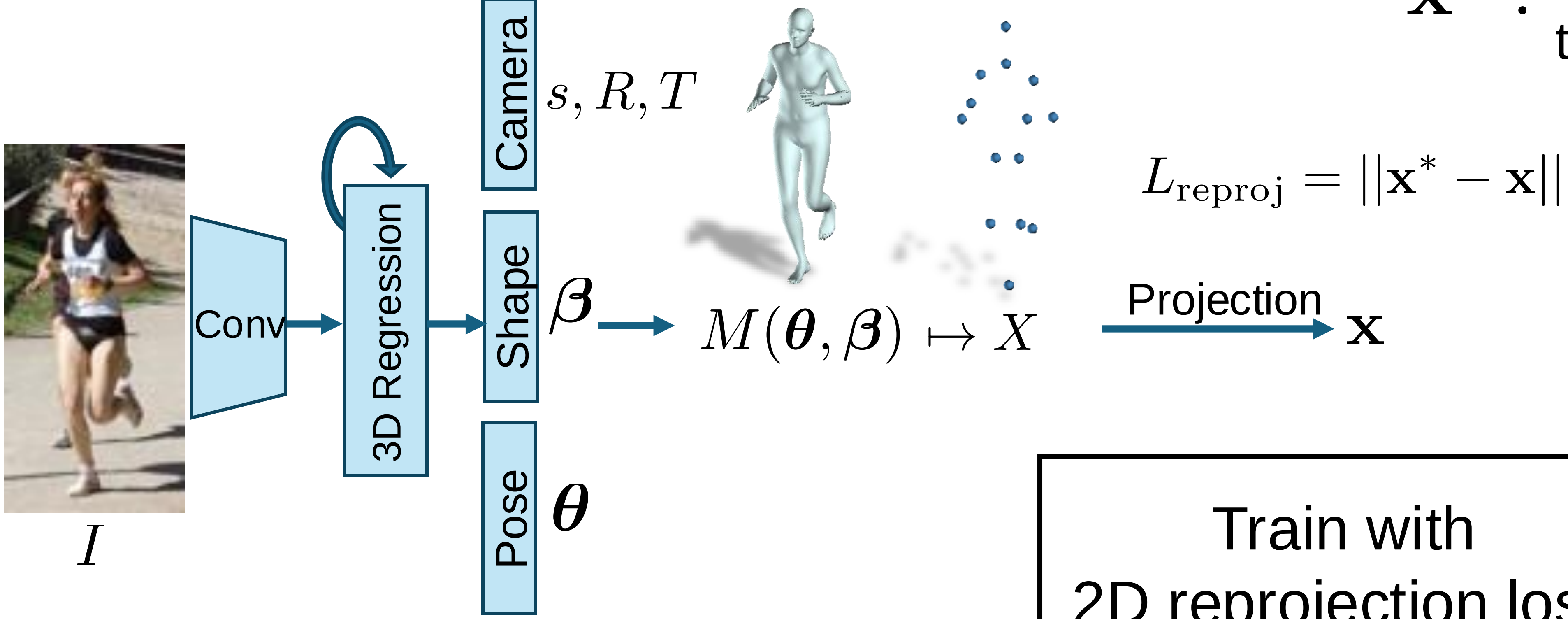


No ground truth for images in the wild!

Kanazawa et al. CVPR 2018

Overview: Human Mesh Recovery (HMR)

$\mathbf{x}^*$: 2D ground truth joints



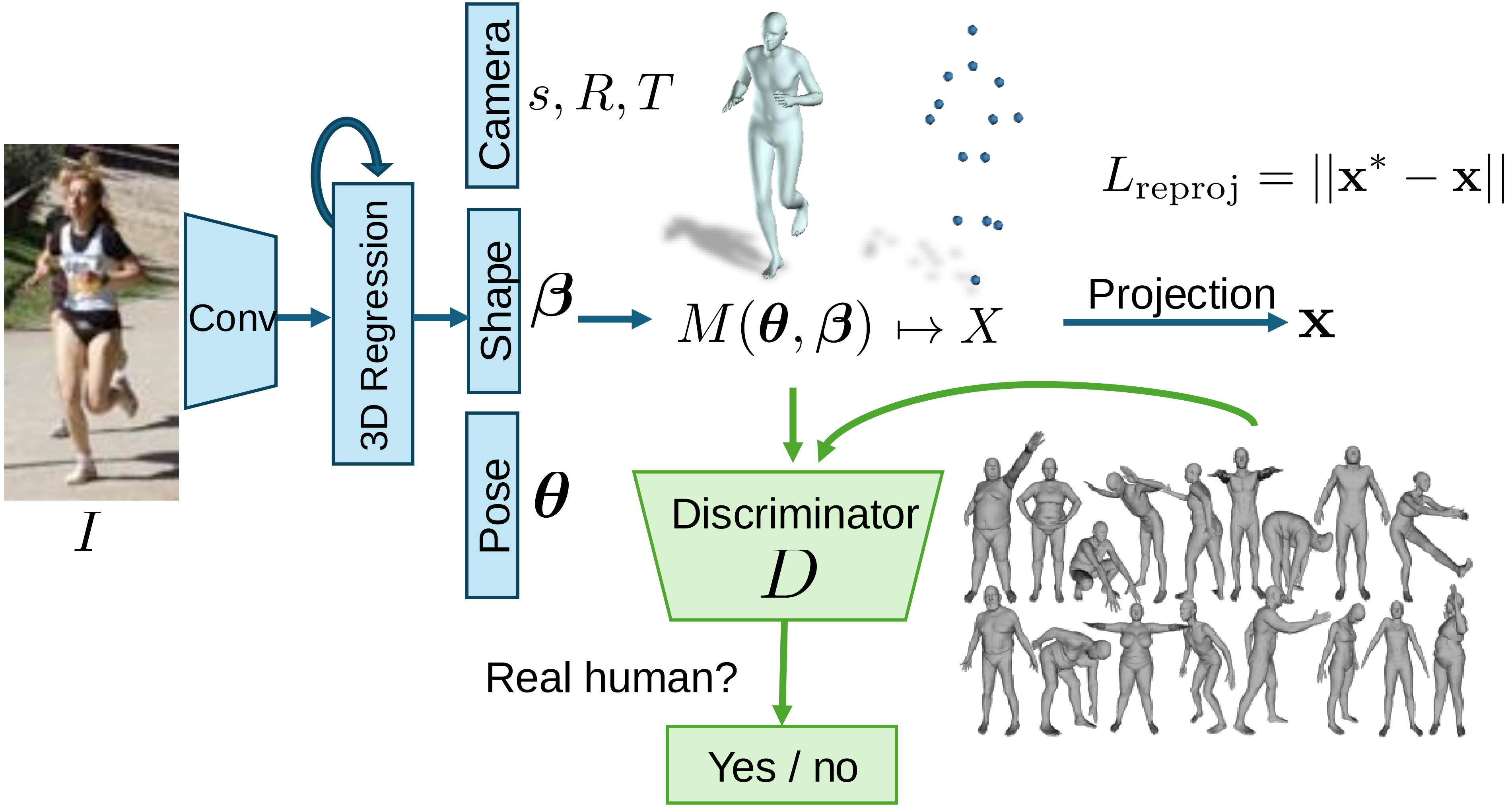
Train with 2D reprojection loss

Kanazawa et al. CVPR 2018

Without any 2D-to-3D supervision...

More monsters from training

Overview: Human Mesh Recovery (HMR)



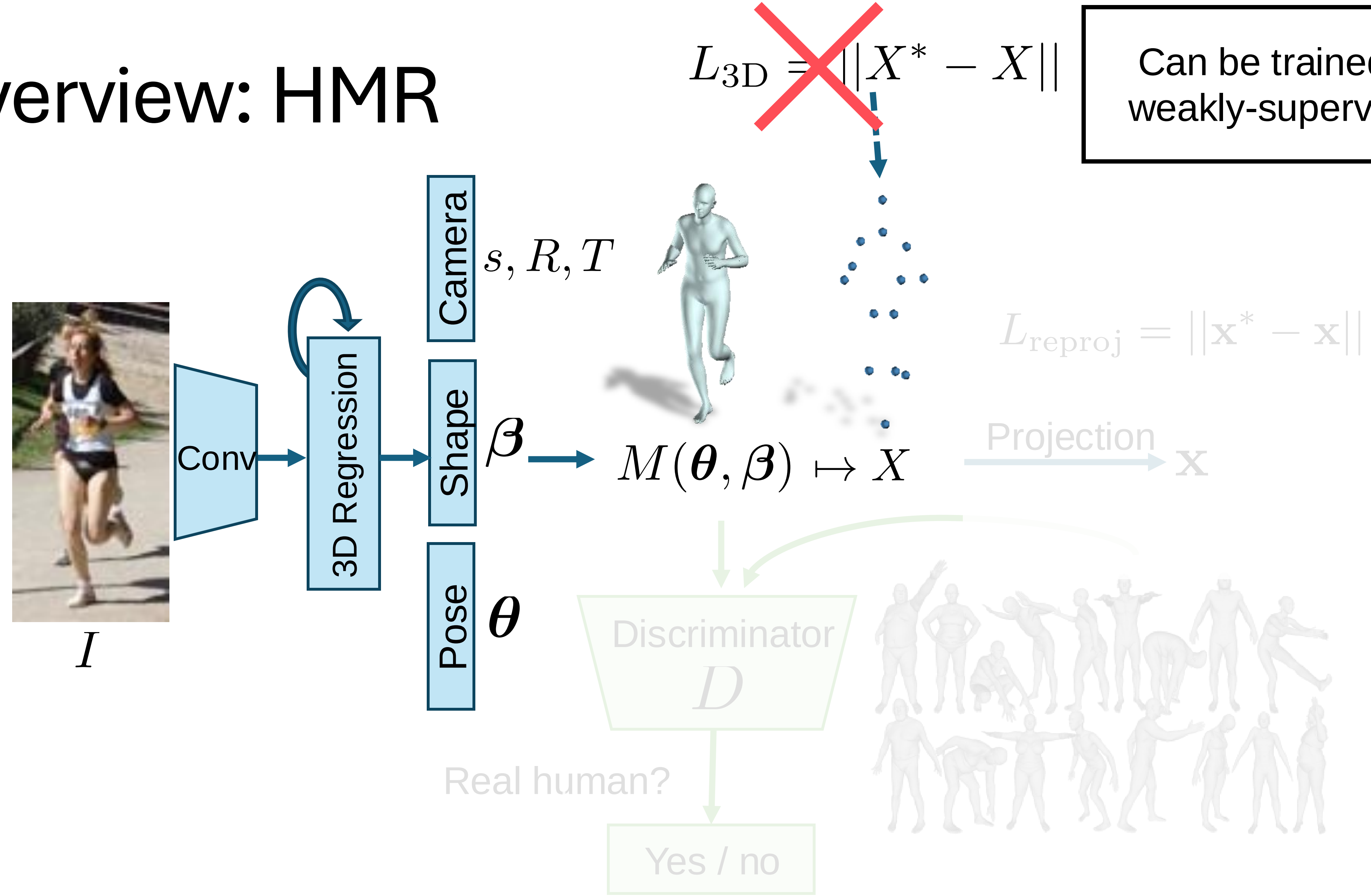
Kanazawa et al. CVPR 2018

Training Data

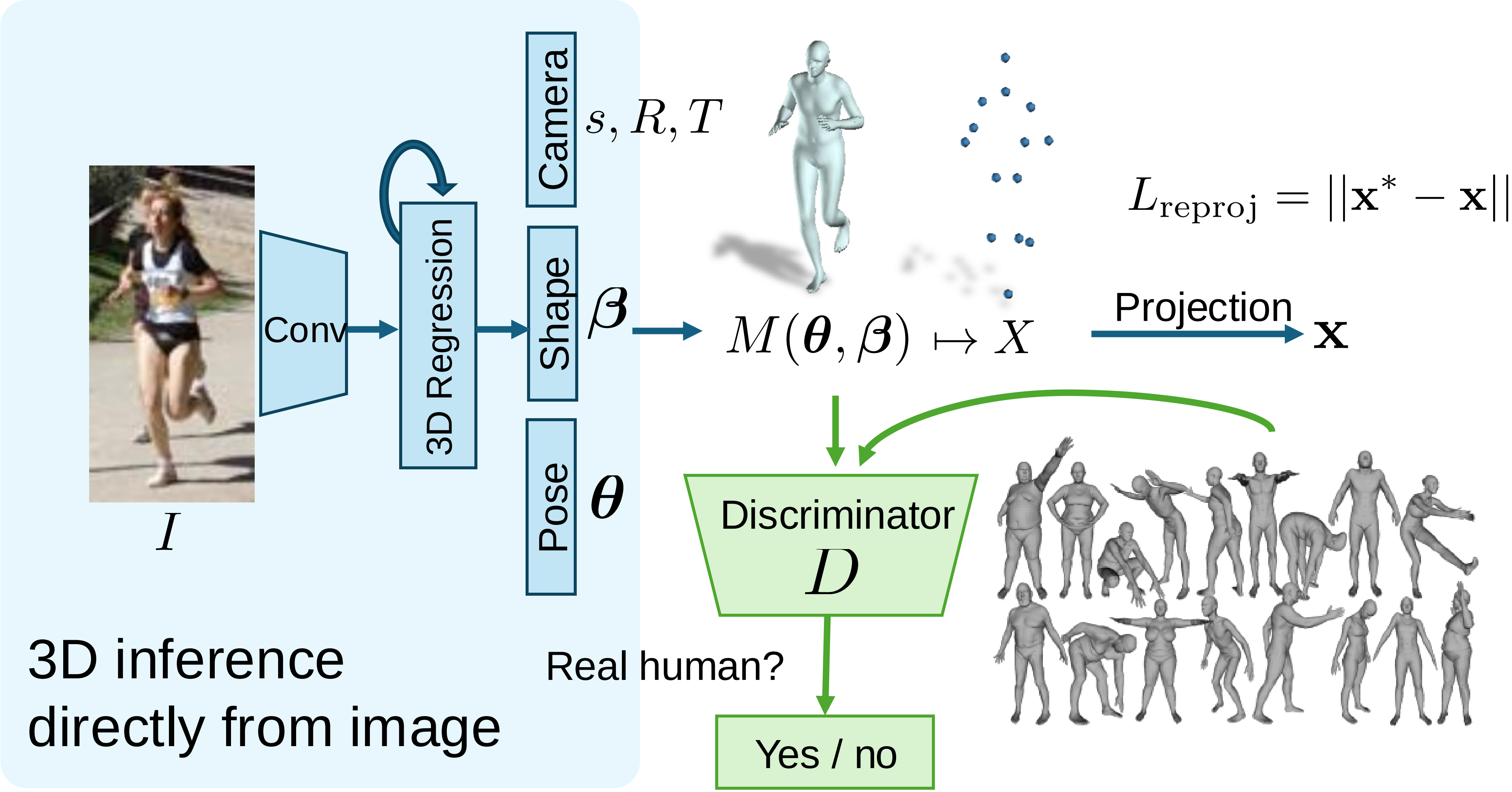
		In-the-wild	3D	2D
Human3.6M [Ionescu et al. PAMI’14]		✗	✓	✓
MS COCO [Lin et al. ECCV ‘14]		✓	✗	✓

Overview: HMR

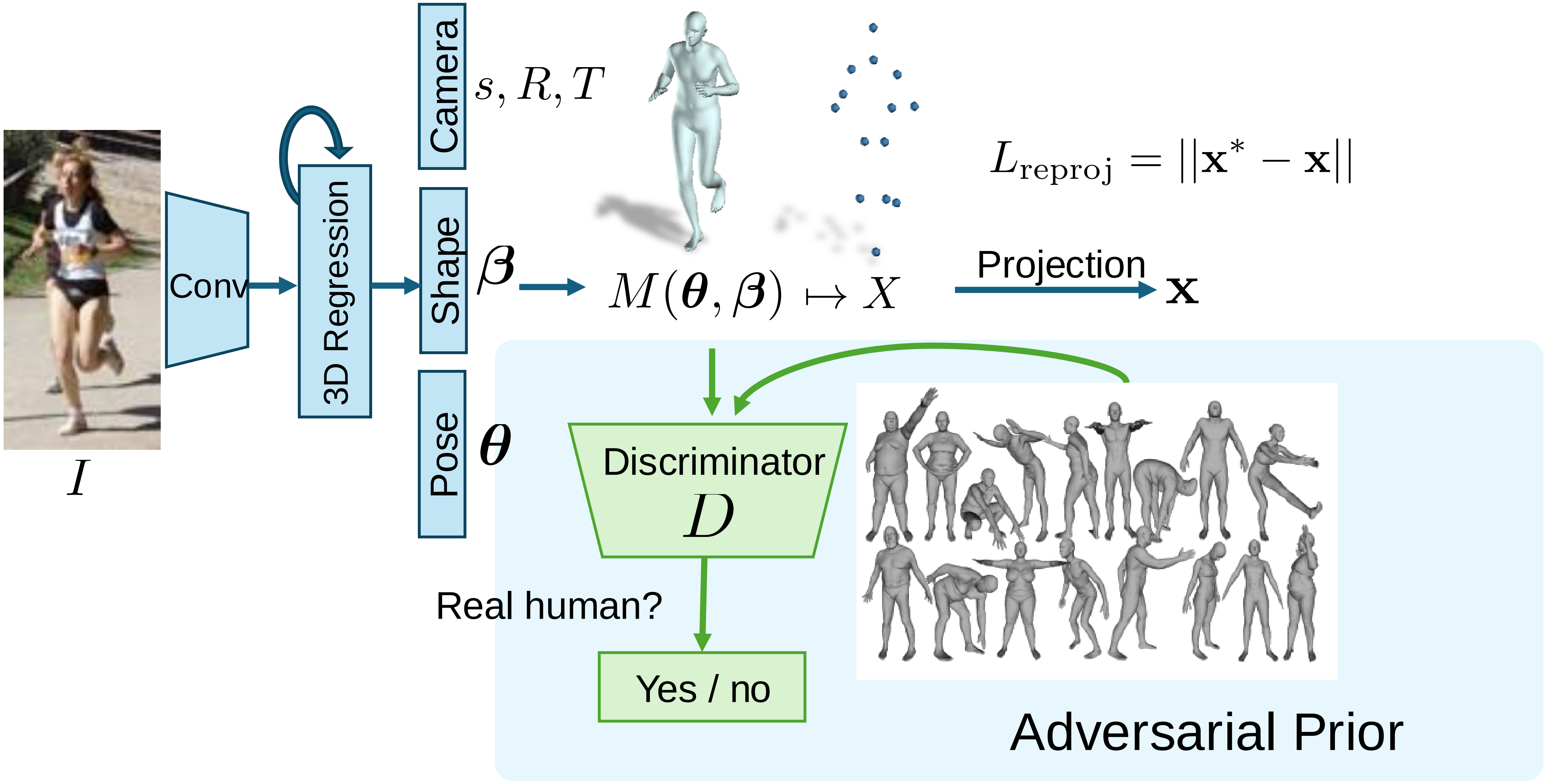
Can be trained in a fully weakly-supervised mode



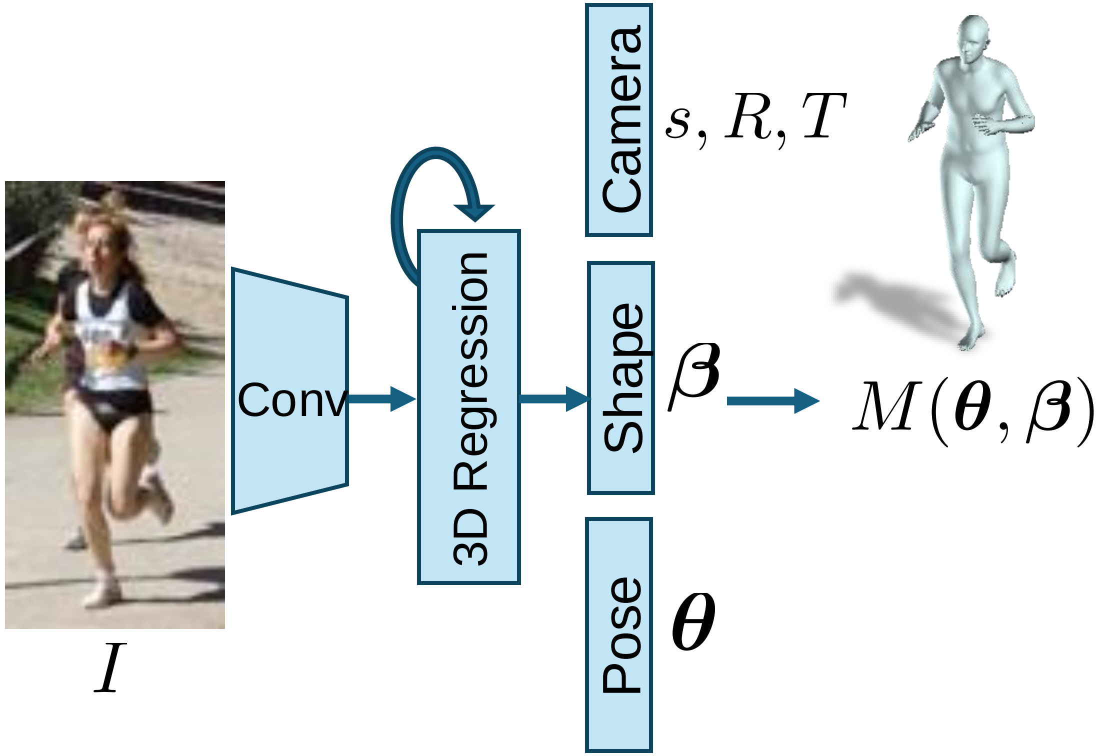
Overview: Human Mesh Recovery (HMR)



Overview: Human Mesh Recovery (HMR)



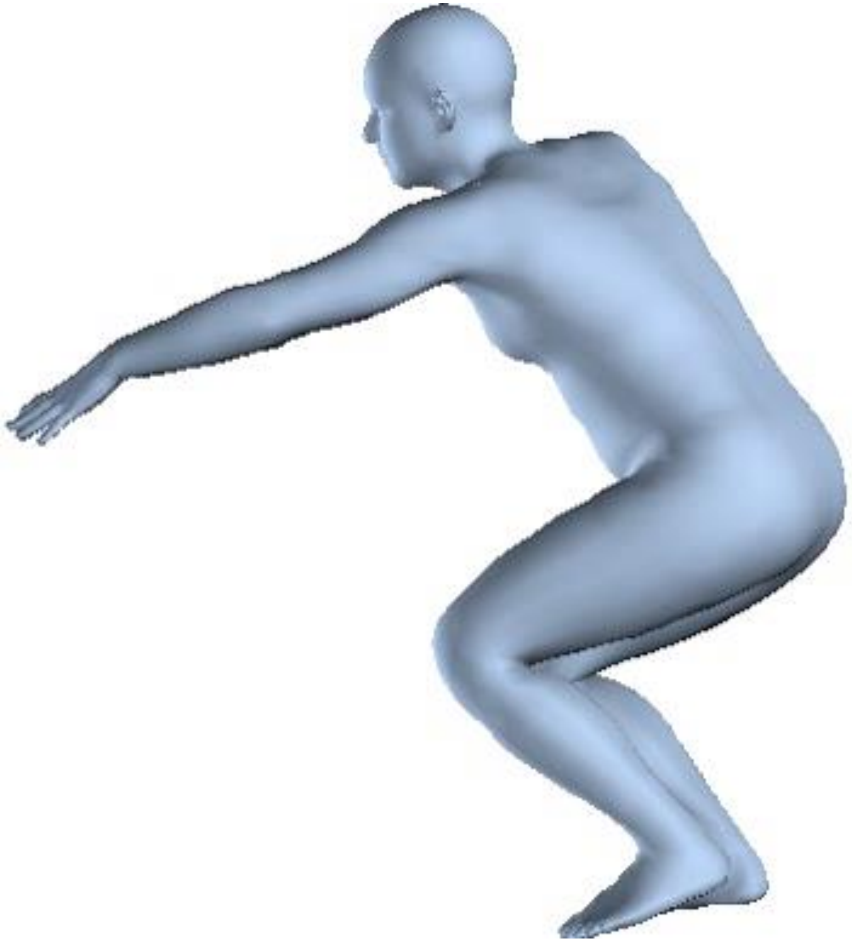
Test time: just feed forward



Benefits of recovering a deformable model

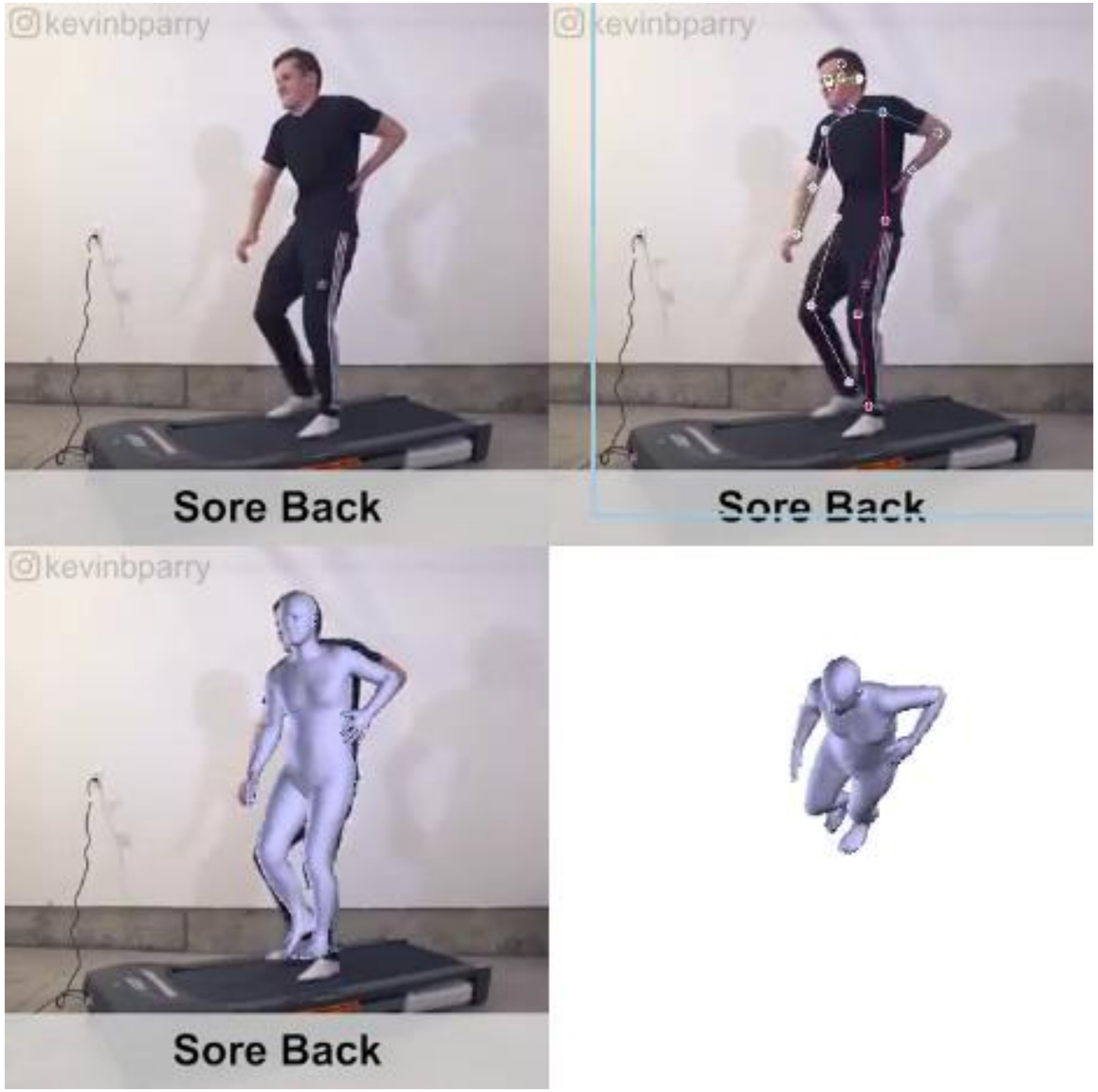
Correspondences across recovered bodies (part segmentations)

Amodal/holistic prediction

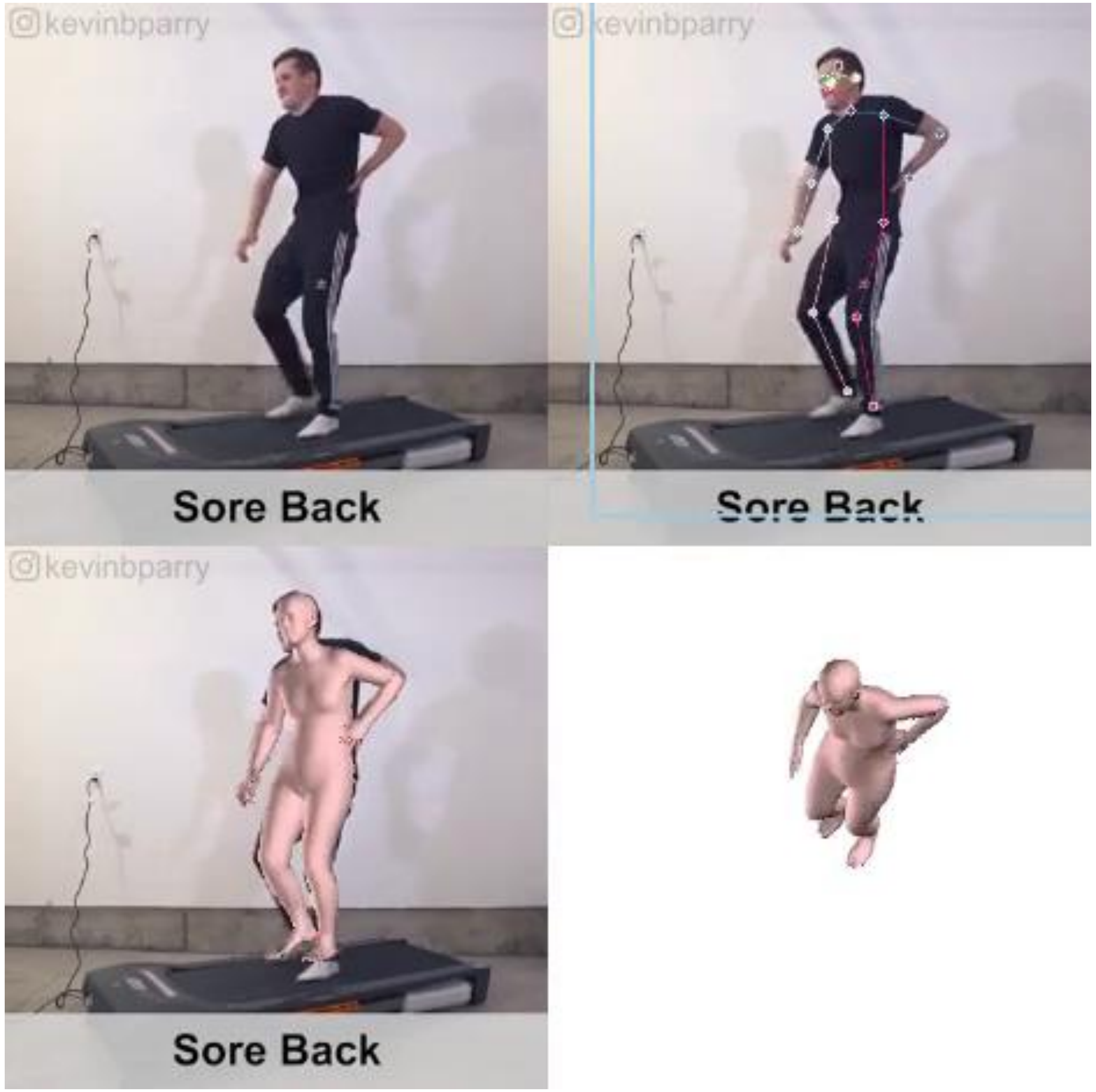


Prediction on occluded body parts

Qualitative results on COCO w/occlusion + clutter



model with full 3D supervision



model **without paired 3D supervision**

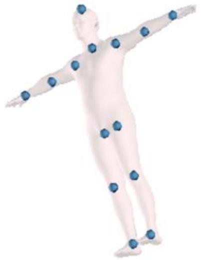

PEOPLE ARE
AWESOME


+ Good per frame performance

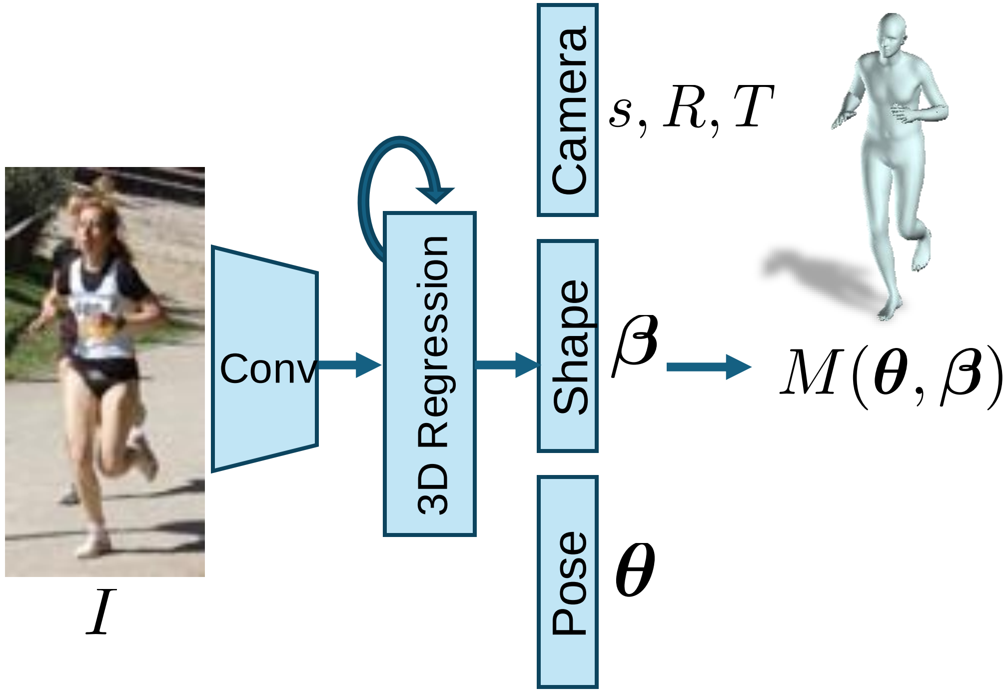
− Lacks temporal coherency


Recap: model based 3D human perception

Iterative optimization in 2016

$\min_{\beta,\theta,\Pi} \Big\|$  $- \Pi($  $) \Big\|_2^2 +$ lots of priors

One-shot inference in 2018

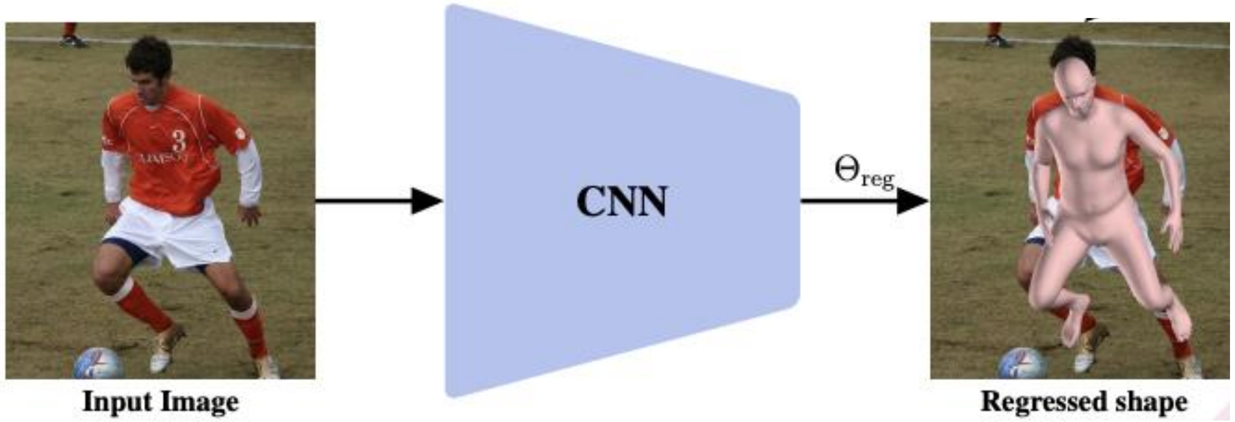


Complementary!
Discuss Pros and Cons

SPIN (SMPL oPtimization IN the loop)

[Kolotouros and Pavlakos et al. ICCV 2019]

SPIN [Kolotouros and Pavlakos et al. ICCV 2019]



SPIN is self-improving

Starting from an initial set of fits, our method can **improve** them.

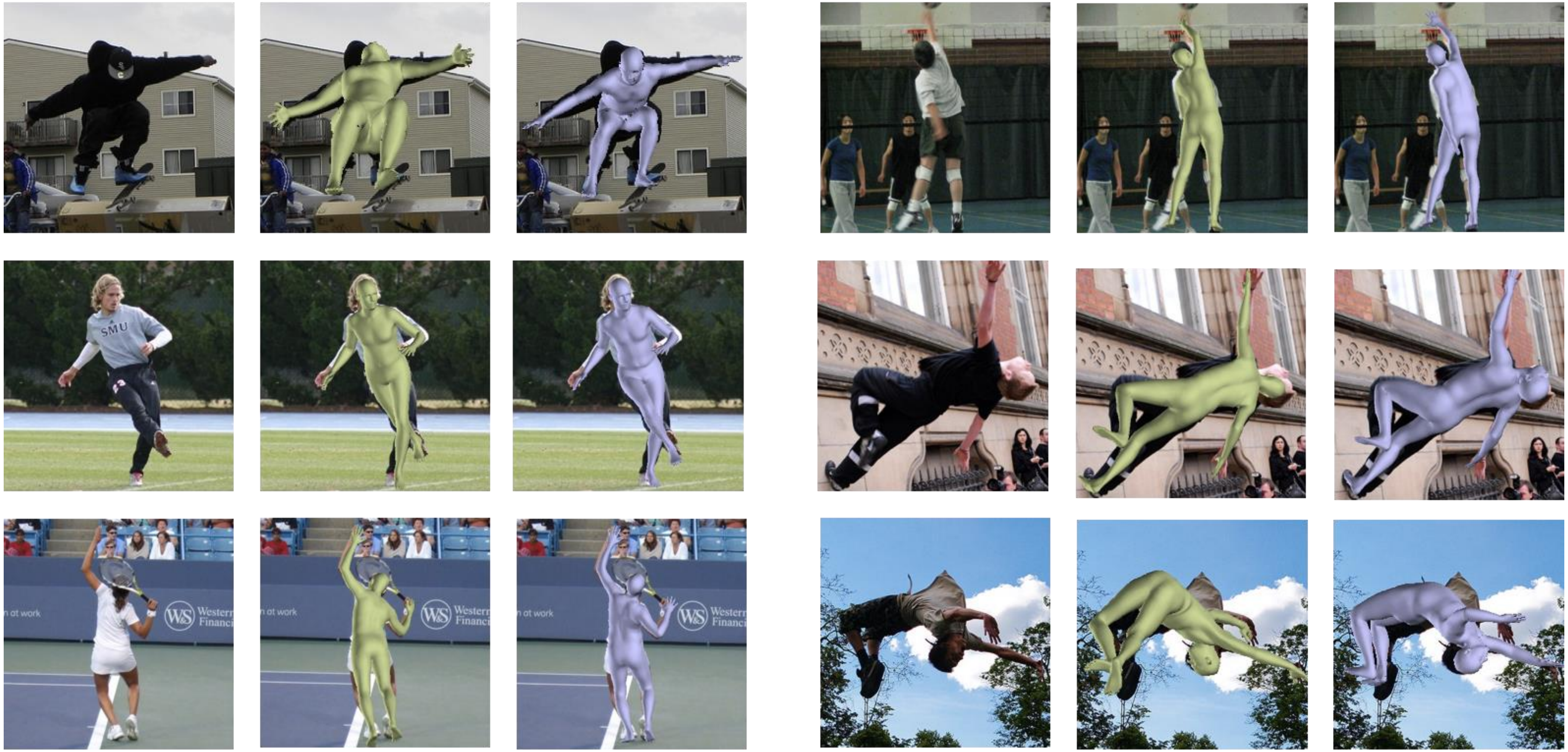
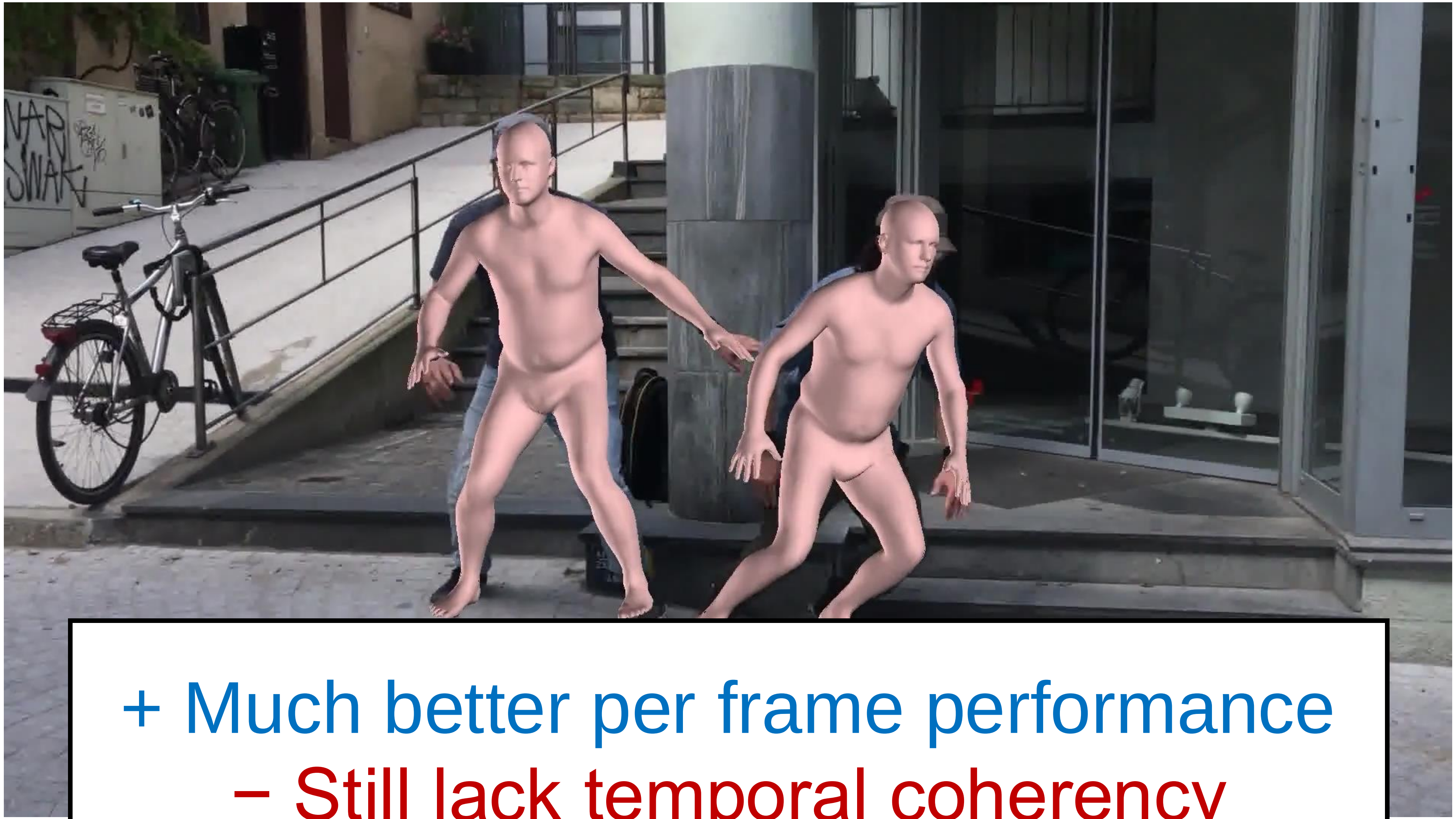


Image | Initial fit | Final fit | Image | Initial fit | Final fit

SPIN results



+ Much better per frame performance

− Still lack temporal coherency

SMPL-X model

SMPL-X estimated independently on each frame

Model fitting

Pavlakos et al.
CVPR 2019

Objective function

$$E(\beta, \theta, \psi) =$$ **Data term** (joints reprojection) + **Priors** (pose, shape, expression, interpenetration)

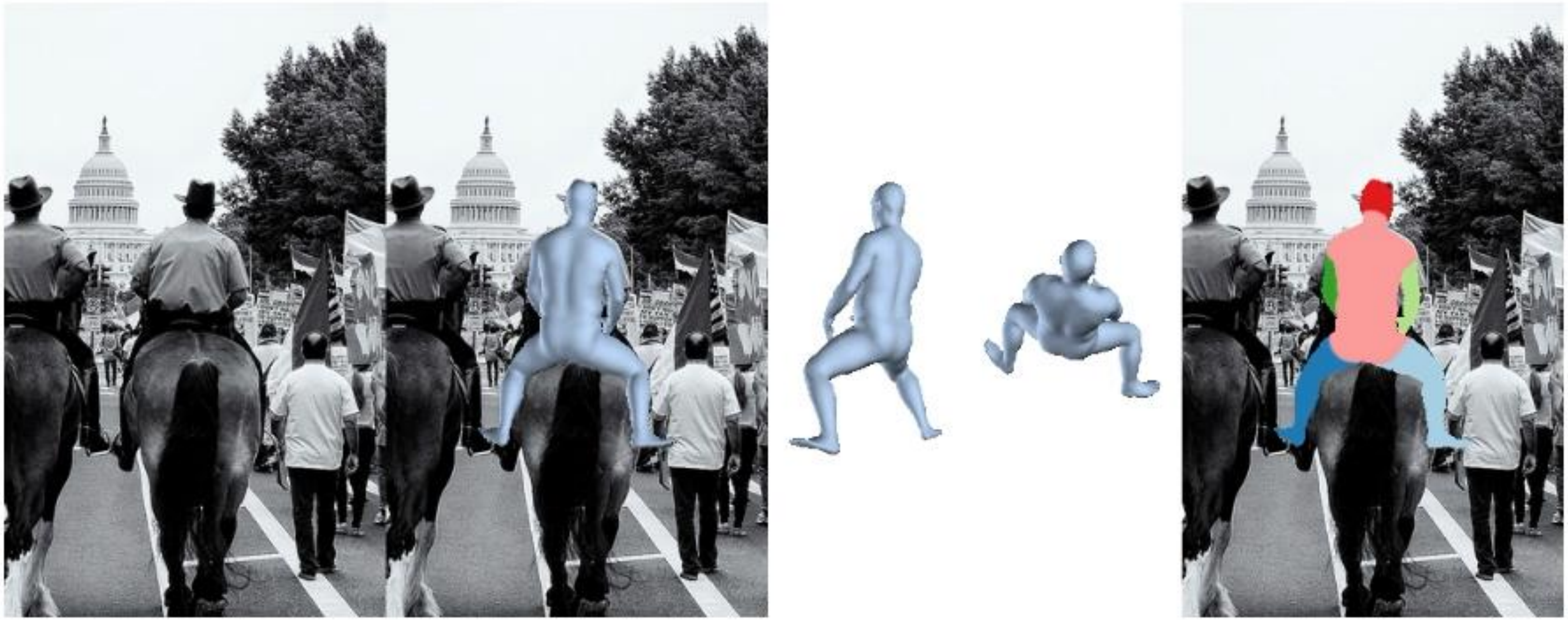
TODO replace with 4D Humans, HaMeR, SLAHMR

Progress on Human Mesh Recovery — from 2018 to 2023

Human Mesh Recovery (HMR)

CVPR 2018

Kanazawa, Black, Jacobs, Malik



Human Mesh Recovery 2.0

ICCV 2023

Goel, Pavlakos, Rajasegaran, Kanazawa*, Malik*, ICCV 2023

Per-frame estimation — no smoothness applied
Color = Identity

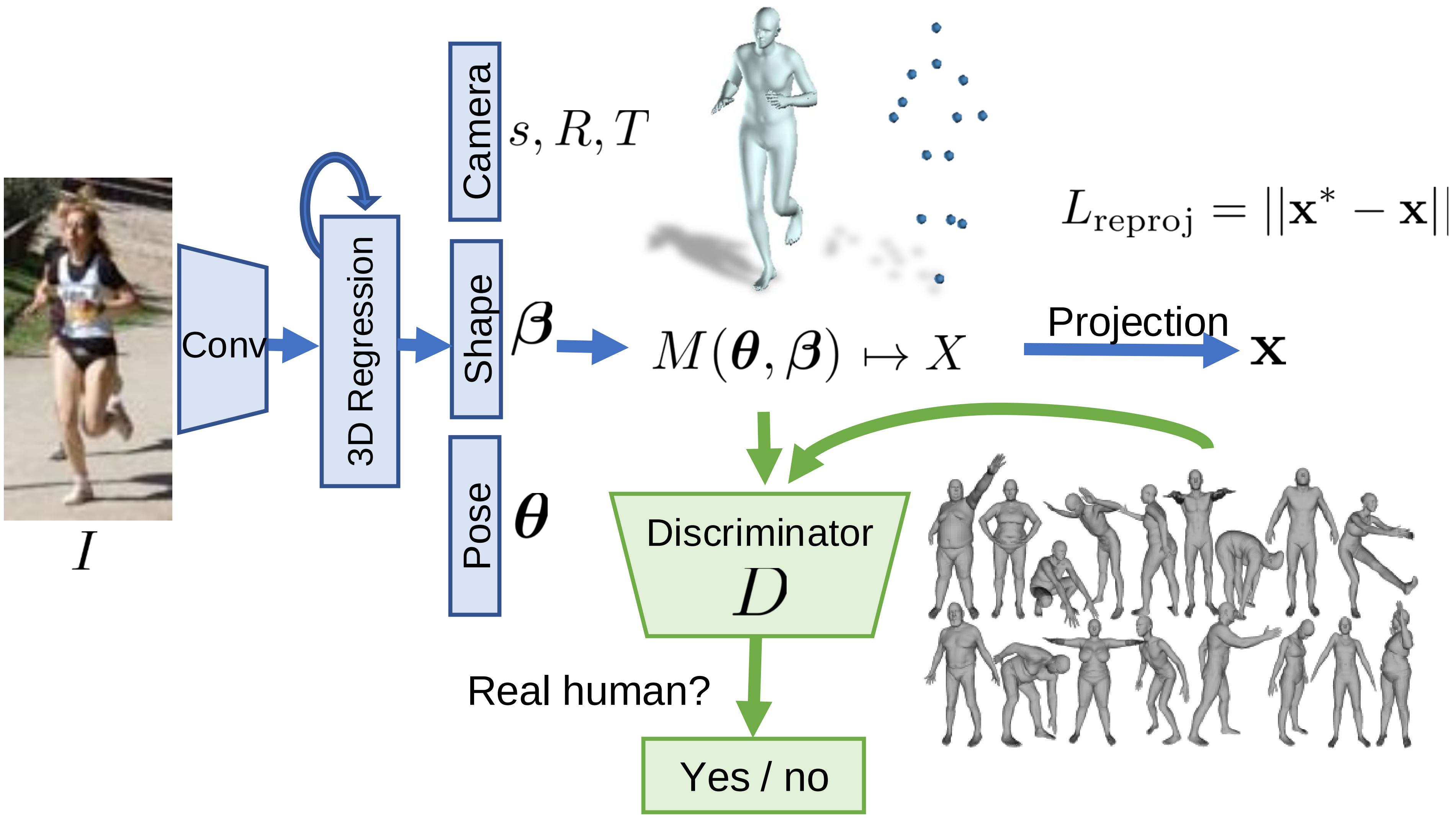
Human Mesh Recovery 2.0

ICCV 2023

Goel, Pavlakos, Rajasegaran, Kanazawa*, Malik*, ICCV 2023

Per-frame estimation — no smoothness applied
Color = Identity

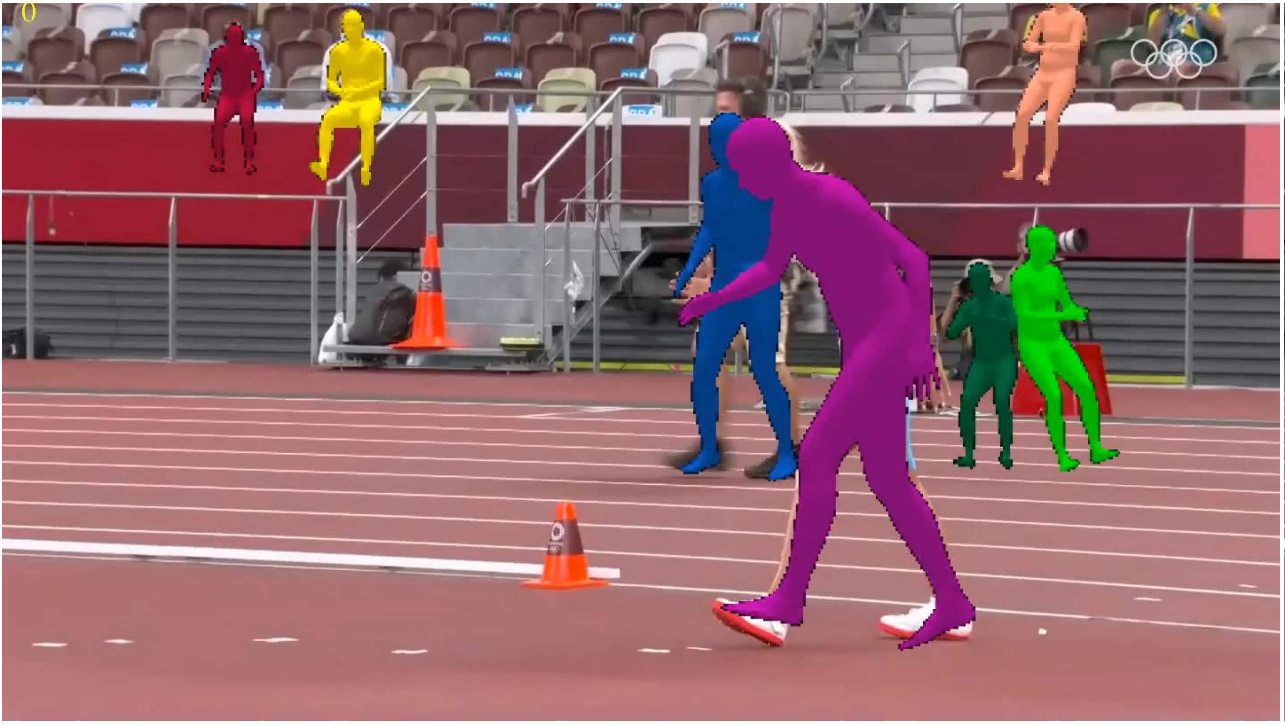
Human Mesh Recovery (HMR) 2018



Kanazawa, Black, Jacobs, Malik, CVPR

Recipe: Big Model and Big Data

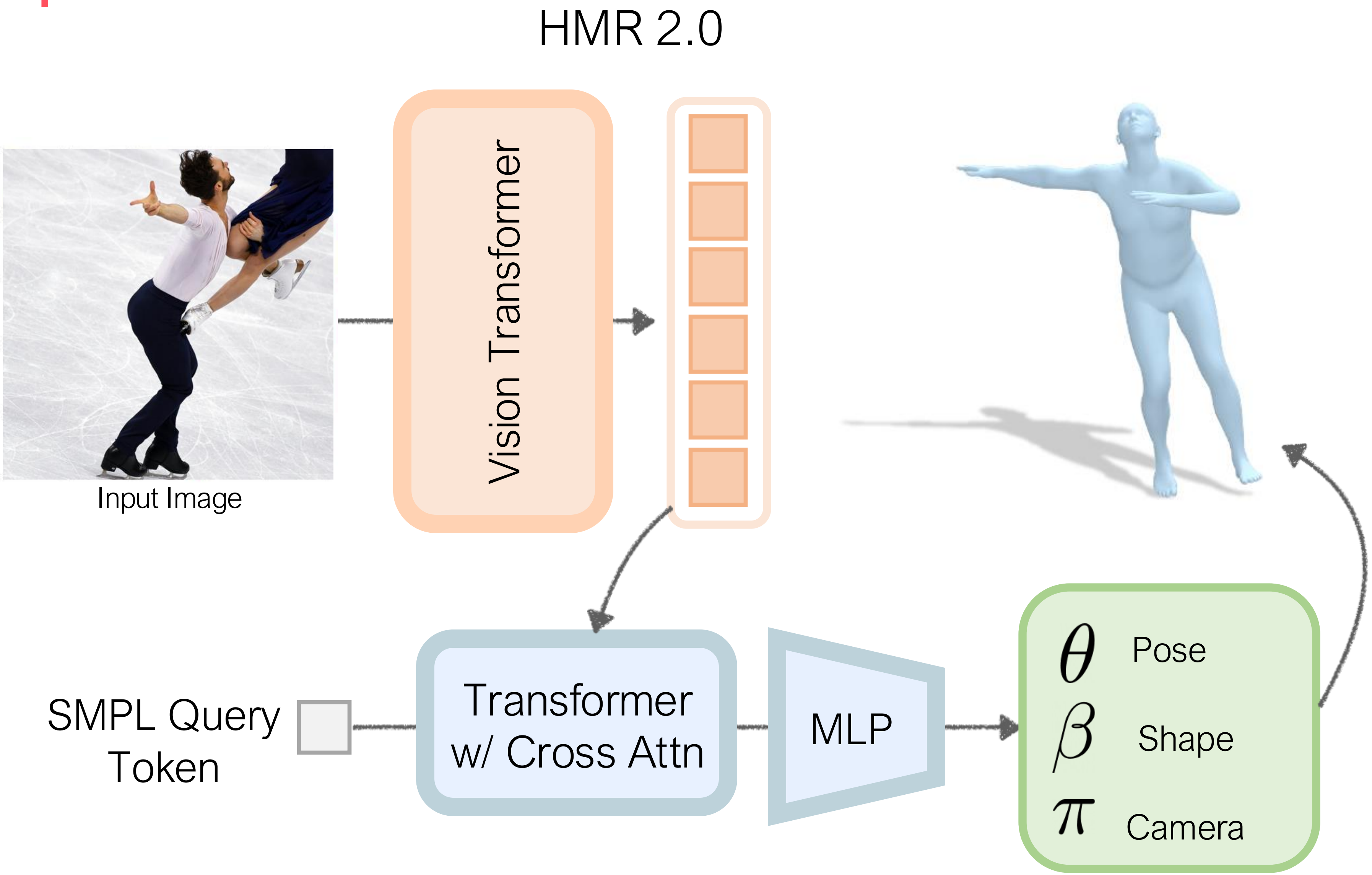
Before



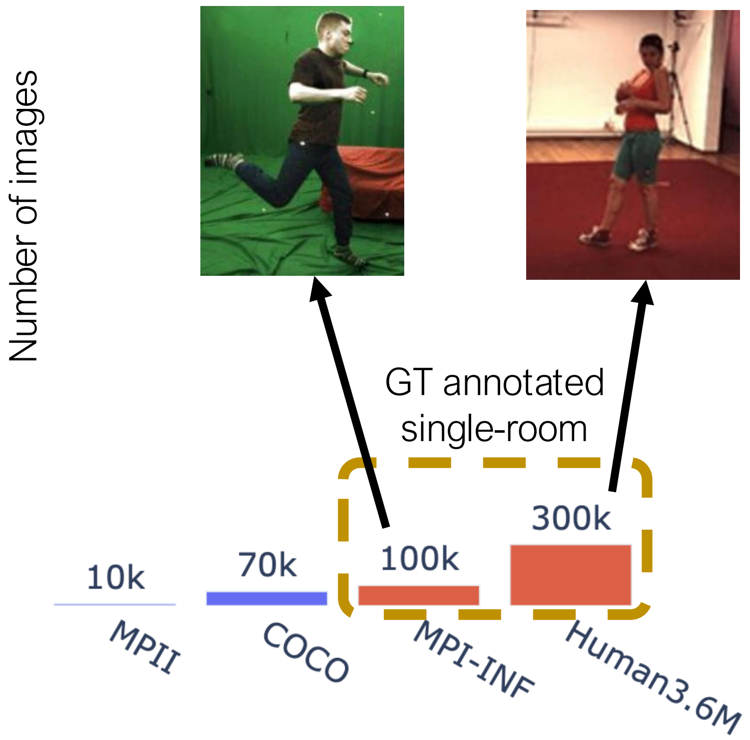
Ours

per-frame estimation - no smoothness applied

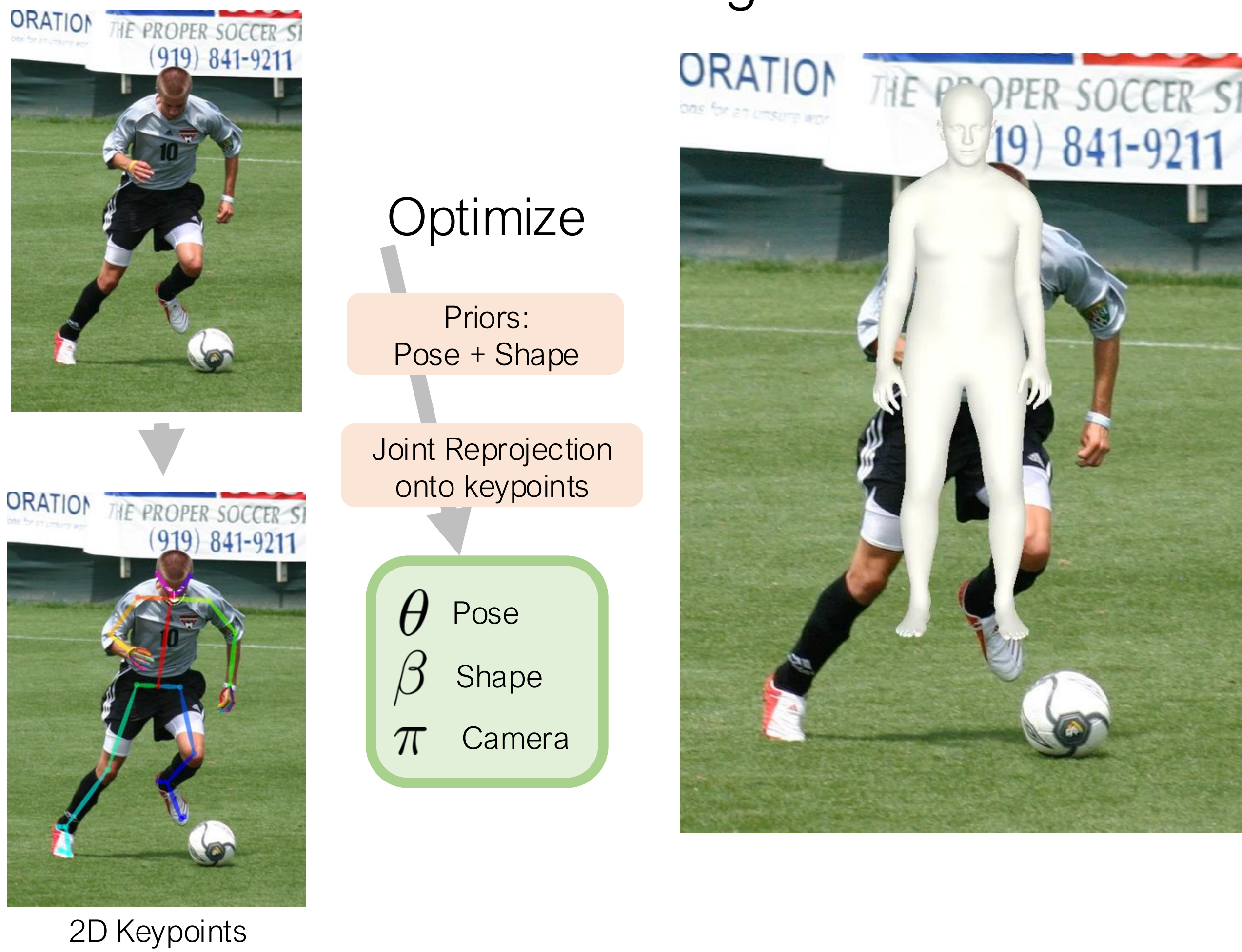
Recipe: Big Model and Big Data



Recipe: Big Model and Big Data



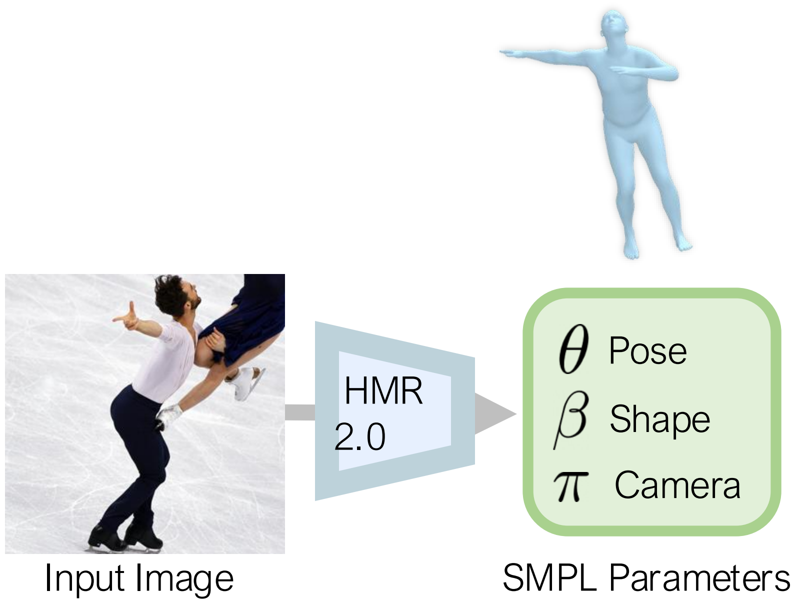
Automatic dataset labelling



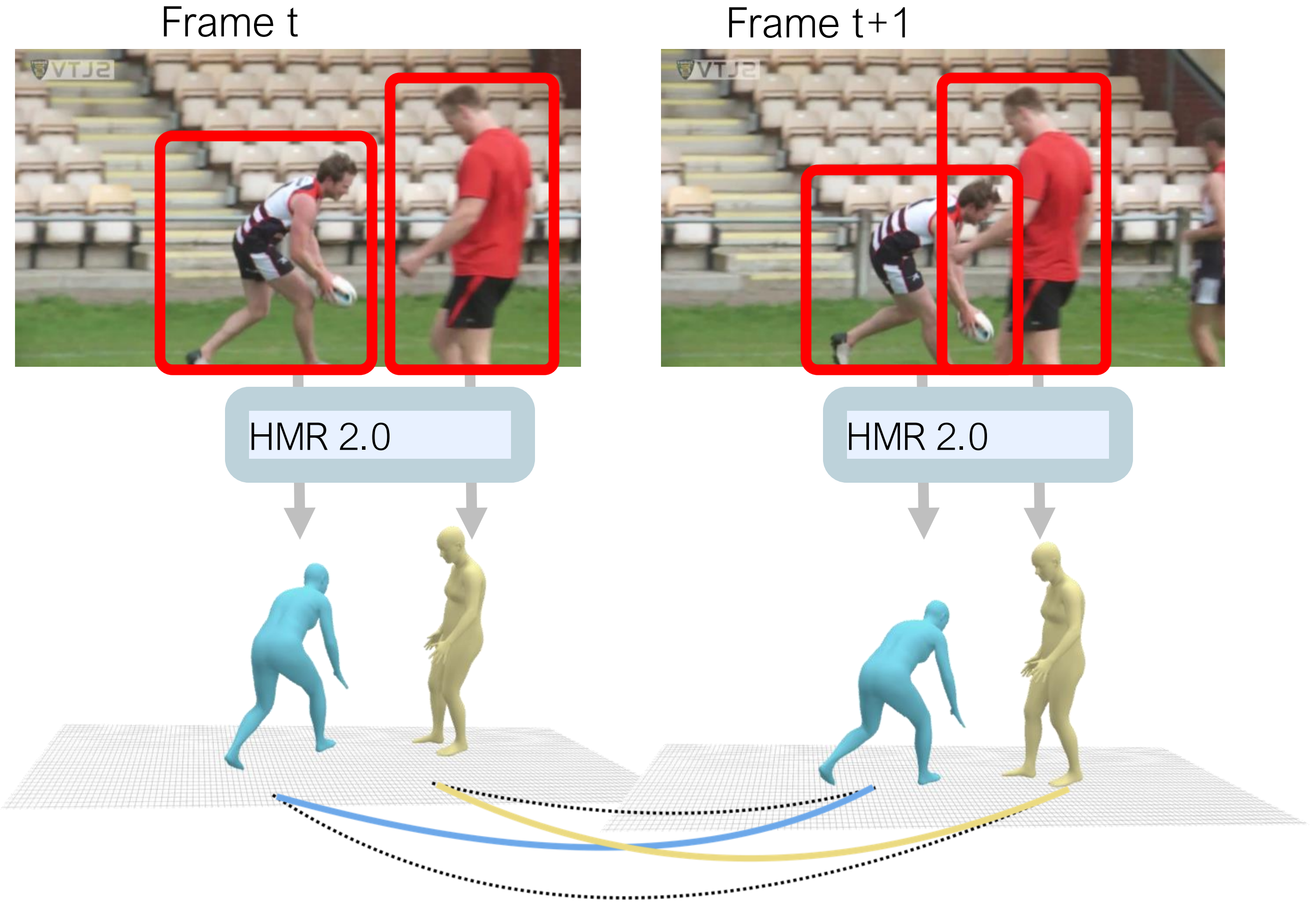
Finally, distill into a network!

4DHumans: HMR2.0 & PHALP++

HMR 2.0

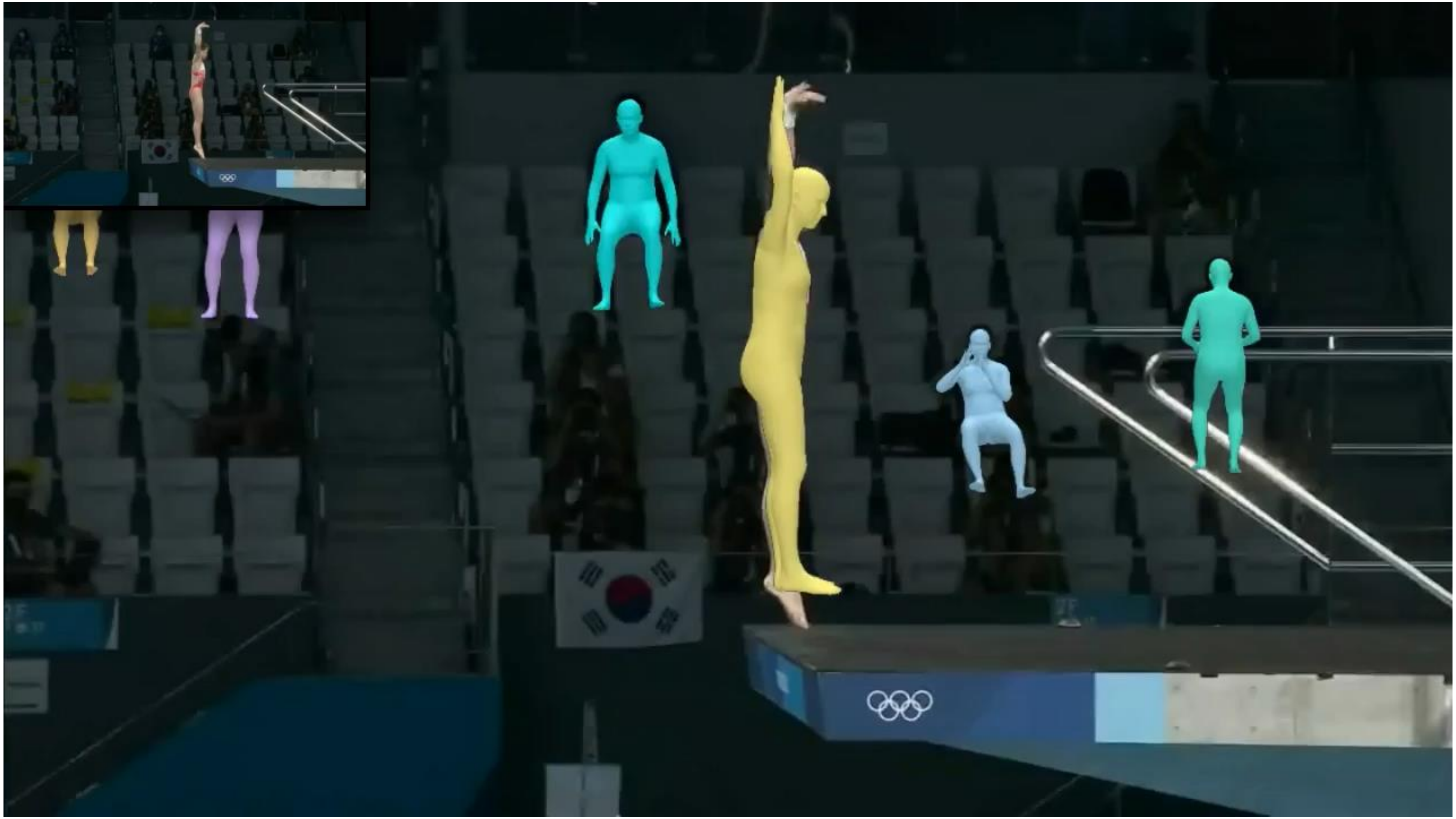


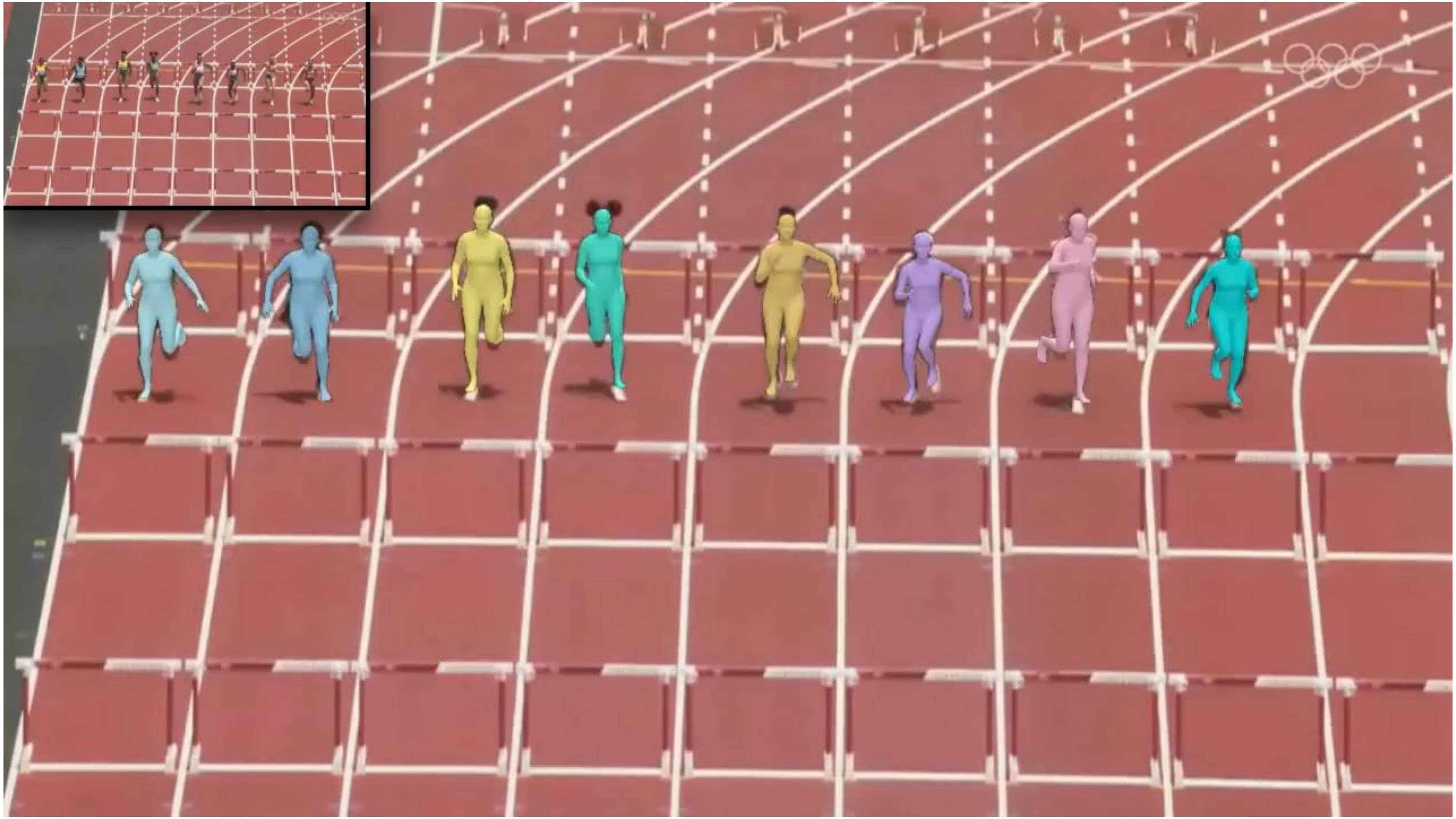
Tracking

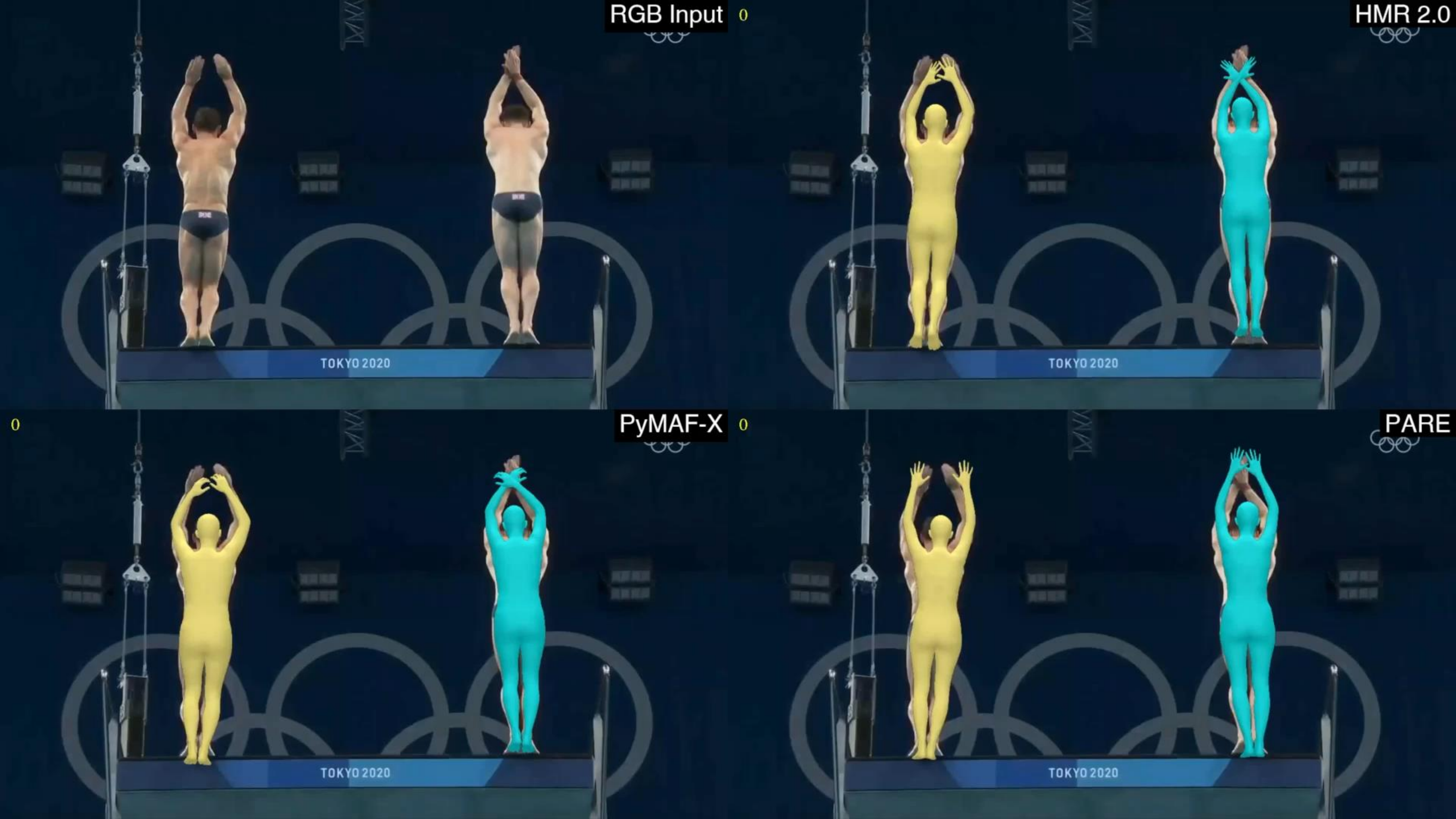


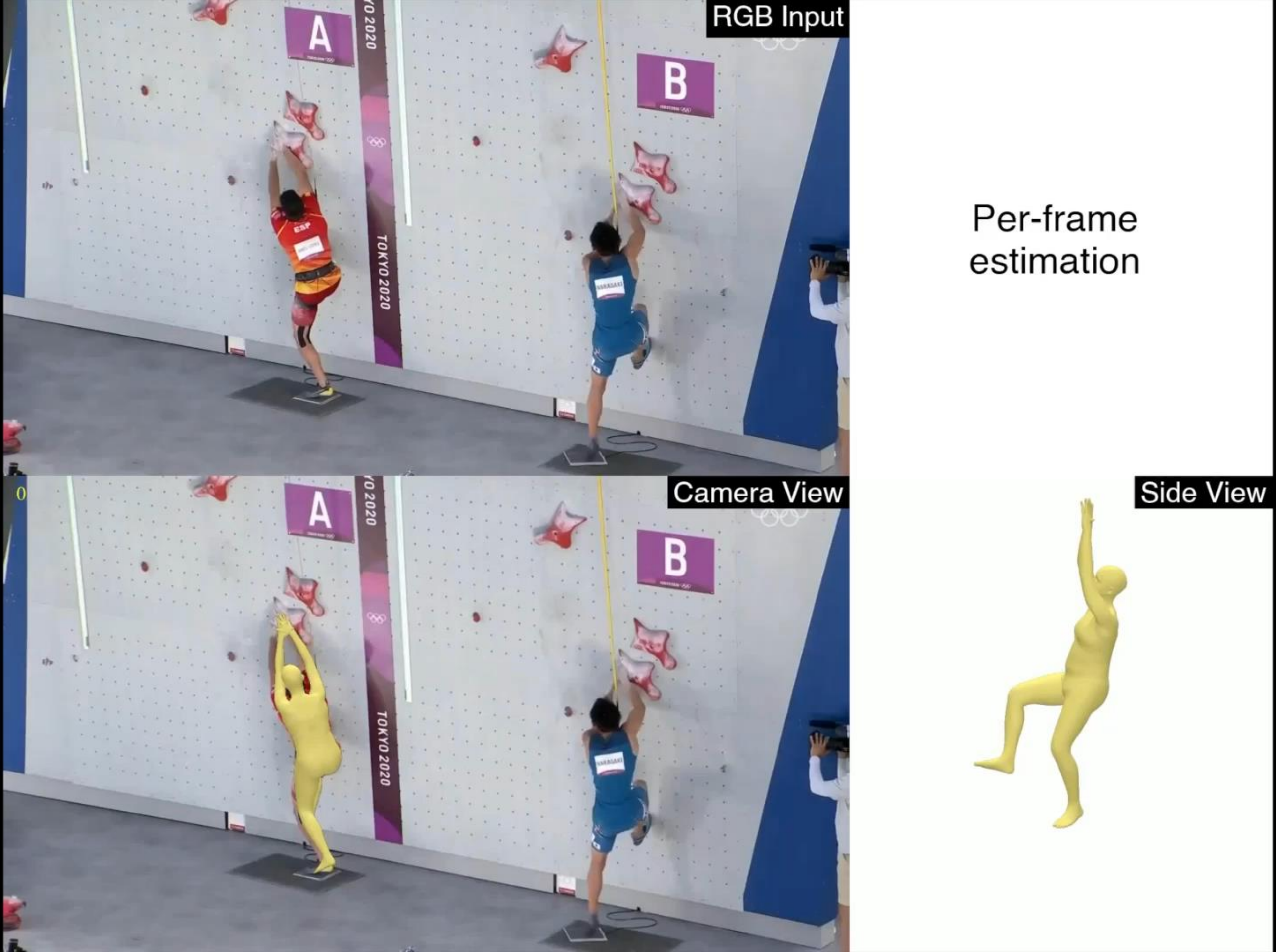
Associate using pose*, location, appearance

Rajasegaran et al. “PHALP”. 2022






RGB Input
0
HMR 2.0
TOKYO 2020
TOKYO 2020
0
PyMAF-X
0
PARE
TOKYO 2020
TOKYO 2020


RGB Input
Per-frame estimation
Camera View
Side View
A
B
TOKYO 2020
0

HaMeR - Hand Mesh Recovery

George Pavlakos

CVPR 2024

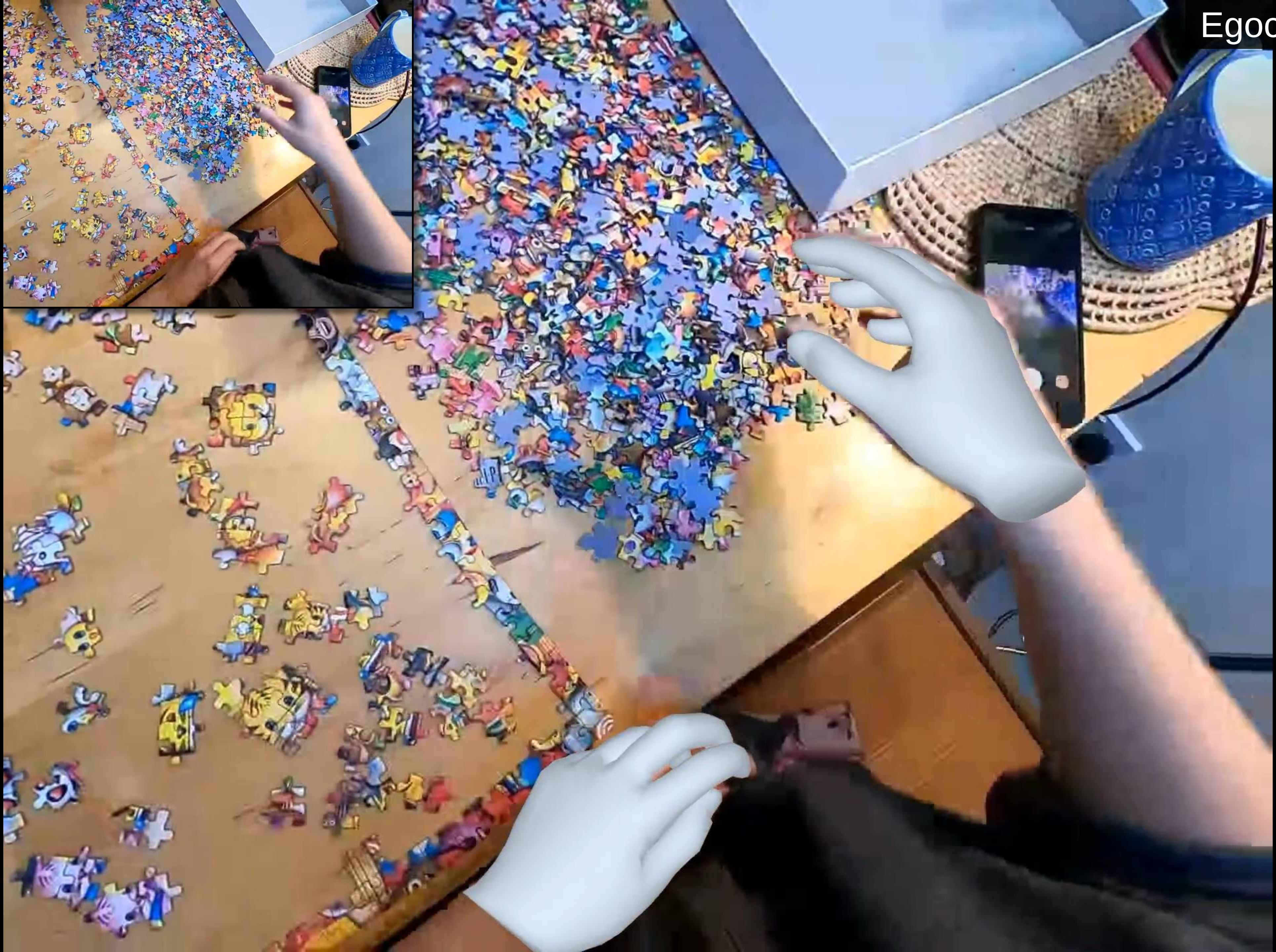

Playing Instruments

Sign Language

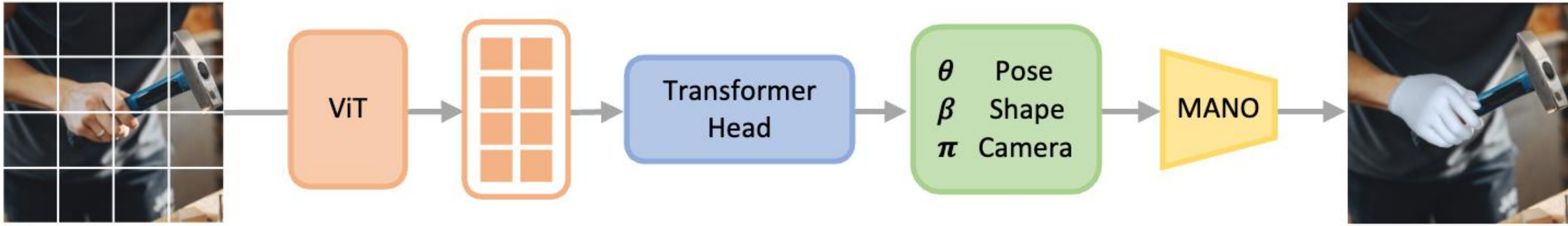

Animation

Hand Gestures

Egocentric Views




Egocentric Views

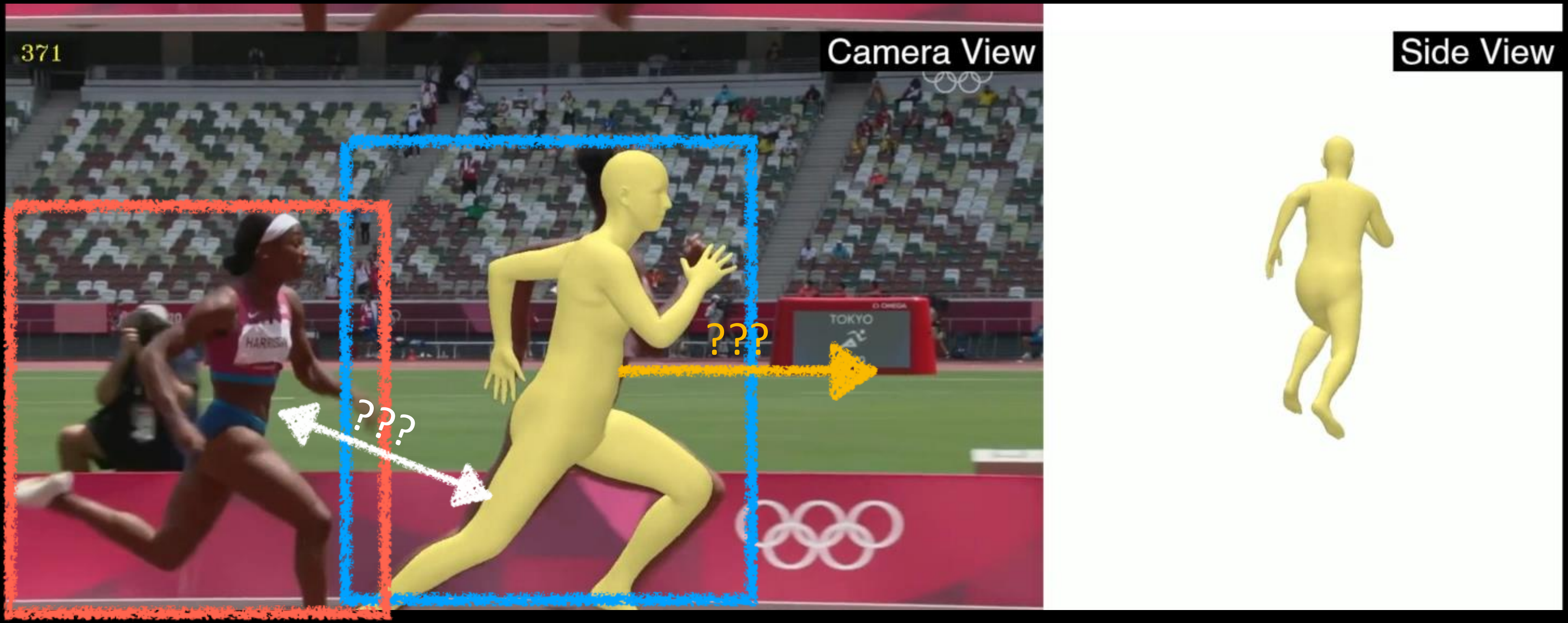

ViT
Transformer Head
θ Pose
β Shape
π Camera
MANO


New Days
(12K)
30%
57%
13%
Ego4D
(23.2K)
VISOR
(5.3K)


RGB Input
HaMeR
Mesh Graphormer
FrankMocap


ITALIAN HAND GESTURE:
VINCENZO'S plate
ITALIAN HAND GESTURE:
VINCENZO'S plate
youtube.com/vincenzosplate
Right hand
Top view
Left hand
Top view

Caveat: Local Pose



Decoupling Human and Camera Motion from Videos in the Wild


Vickie Ye


Georgios Pavlakos


Jitendra Malik

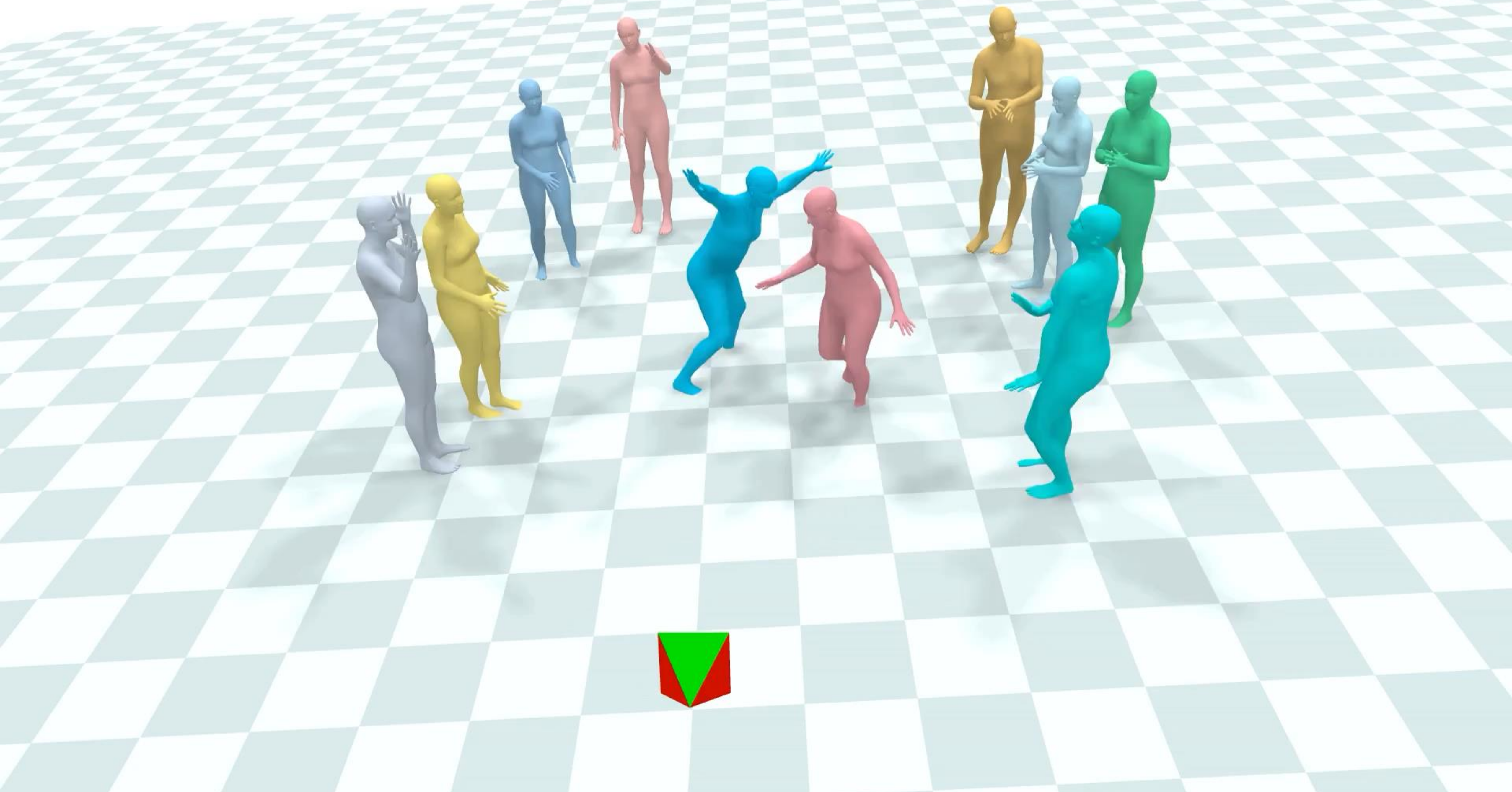

Angjoo Kanazawa

CVPR 2023

UC Berkeley

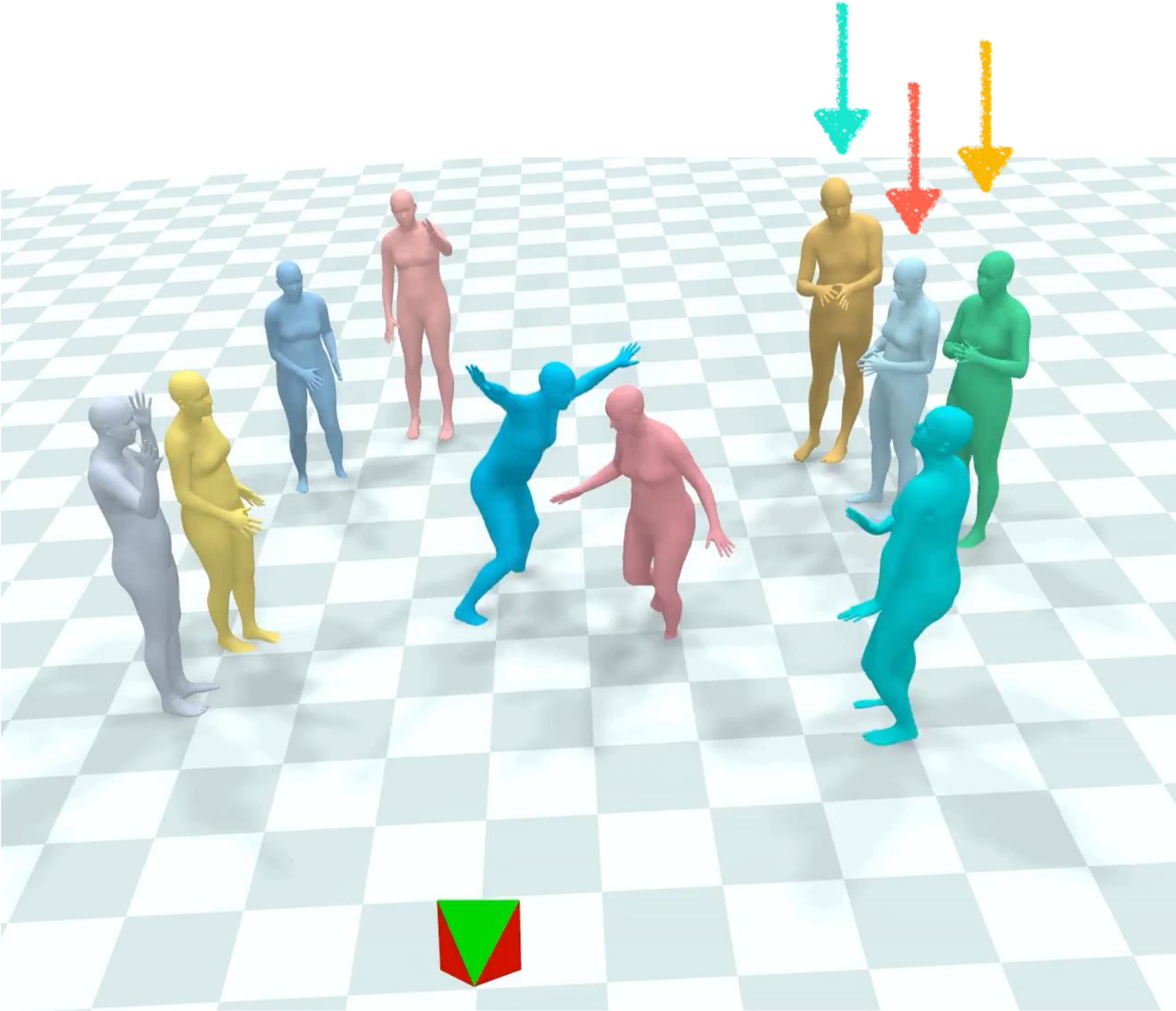

We live in a world that is 3D and dynamic.

Results from SLAHMR!! [Ye et al. CVPR 2023]



SLAHMR: Decoupling Human and Camera Motion from Videos in the Wild. [Ye et al. CVPR 2023]

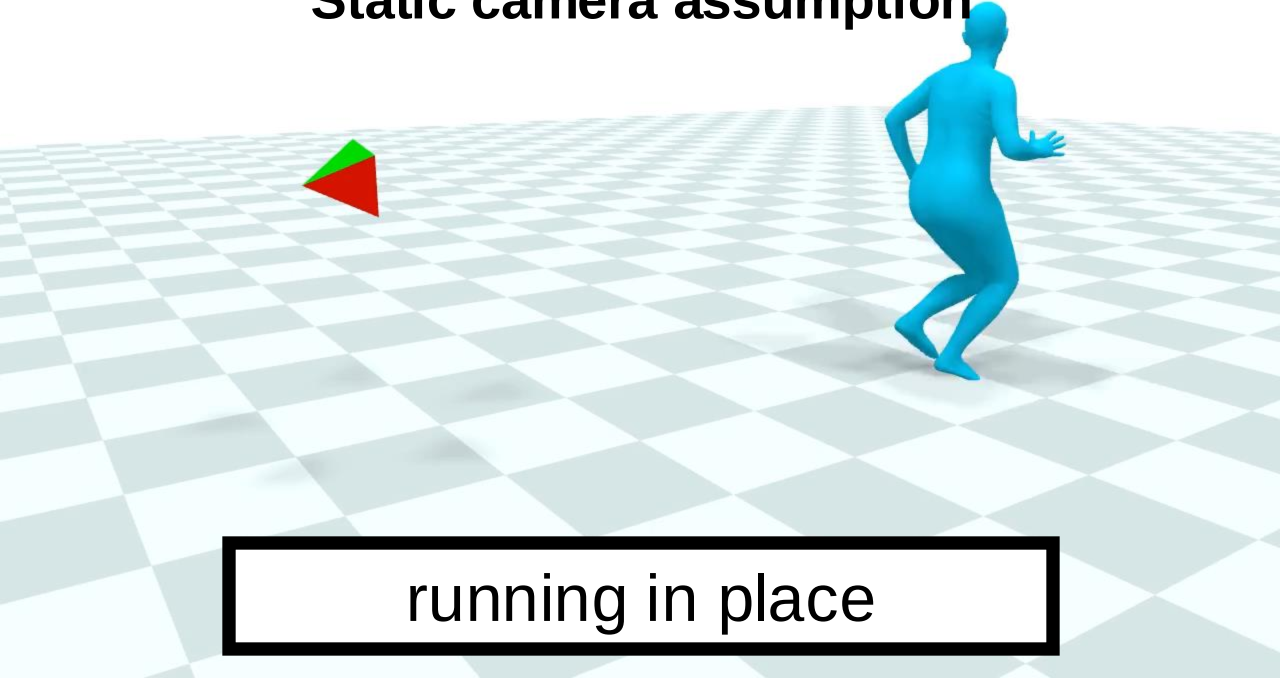
SLAHMR: Decoupling Human and Camera Motion from Videos in the Wild. [Ye et al. CVPR 2023]



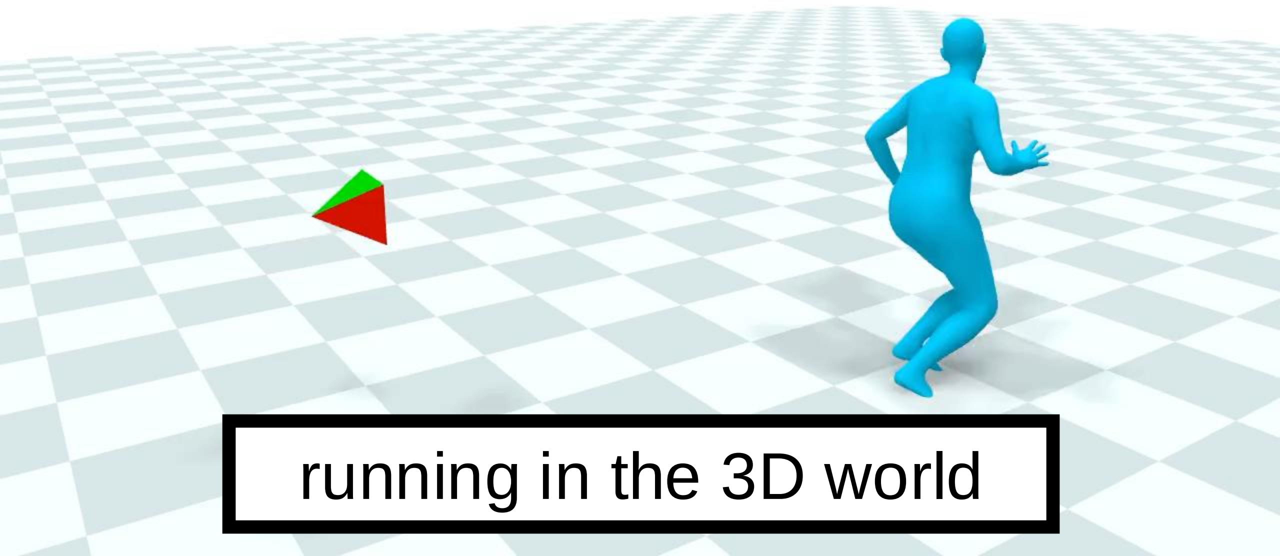
SLAHMR: Decoupling Human and Camera Motion from Videos in the Wild. [Ye et al. CVPR 2023]


WARRIORS
1

Static camera assumption



Modeling camera motion



Input:

$\{{}^{\text{cam}}\mathbf{P}\}$ Tracked People in Unscaled Camera Frame

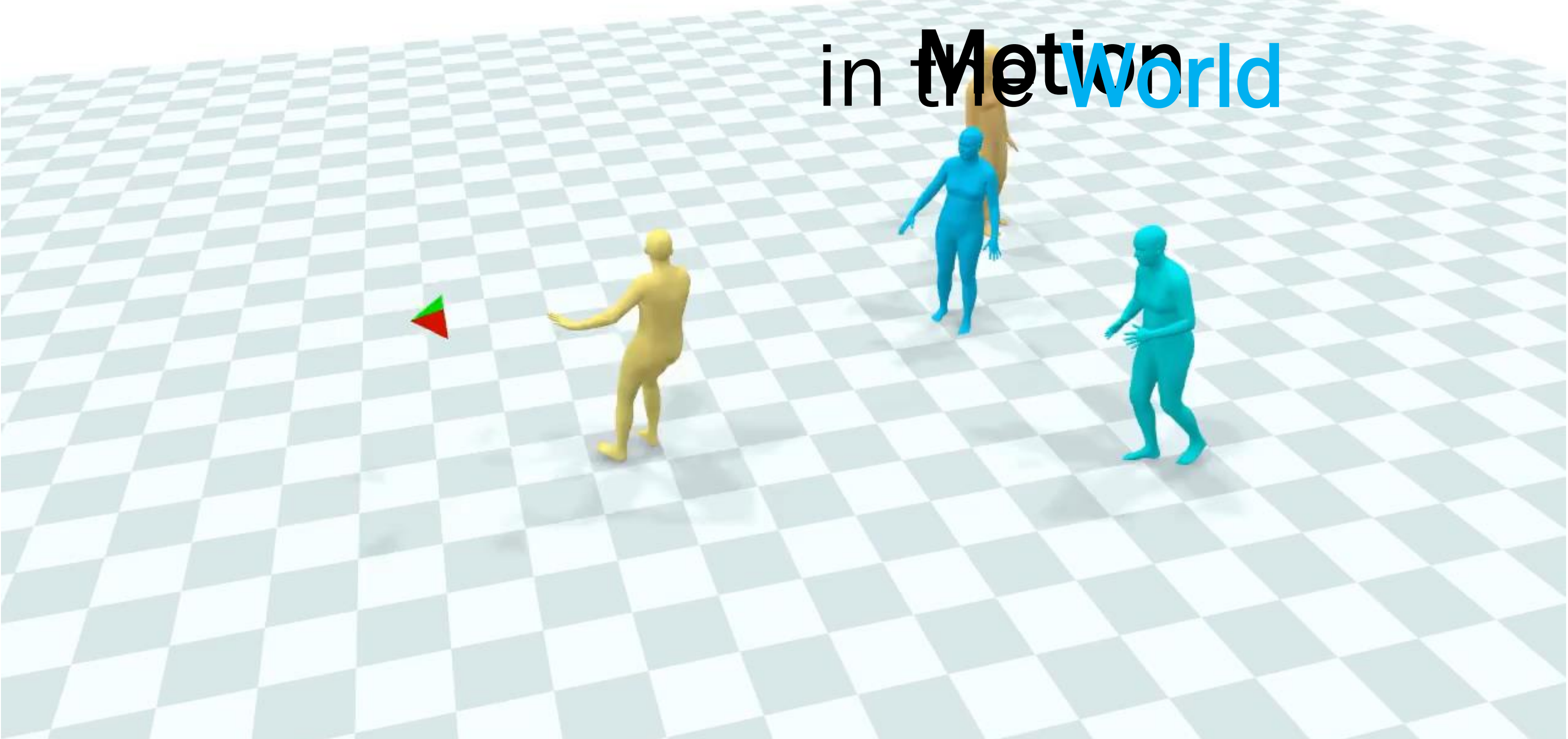
$\{R, T\}$ Camera Motion

Output:

$\{{}^{\text{world}}\mathbf{P}\}$ Tracked People in World Frame

$\{R, \alpha T\}$ Scaled Camera Motion

$\{g\}$ Ground Plane in the World

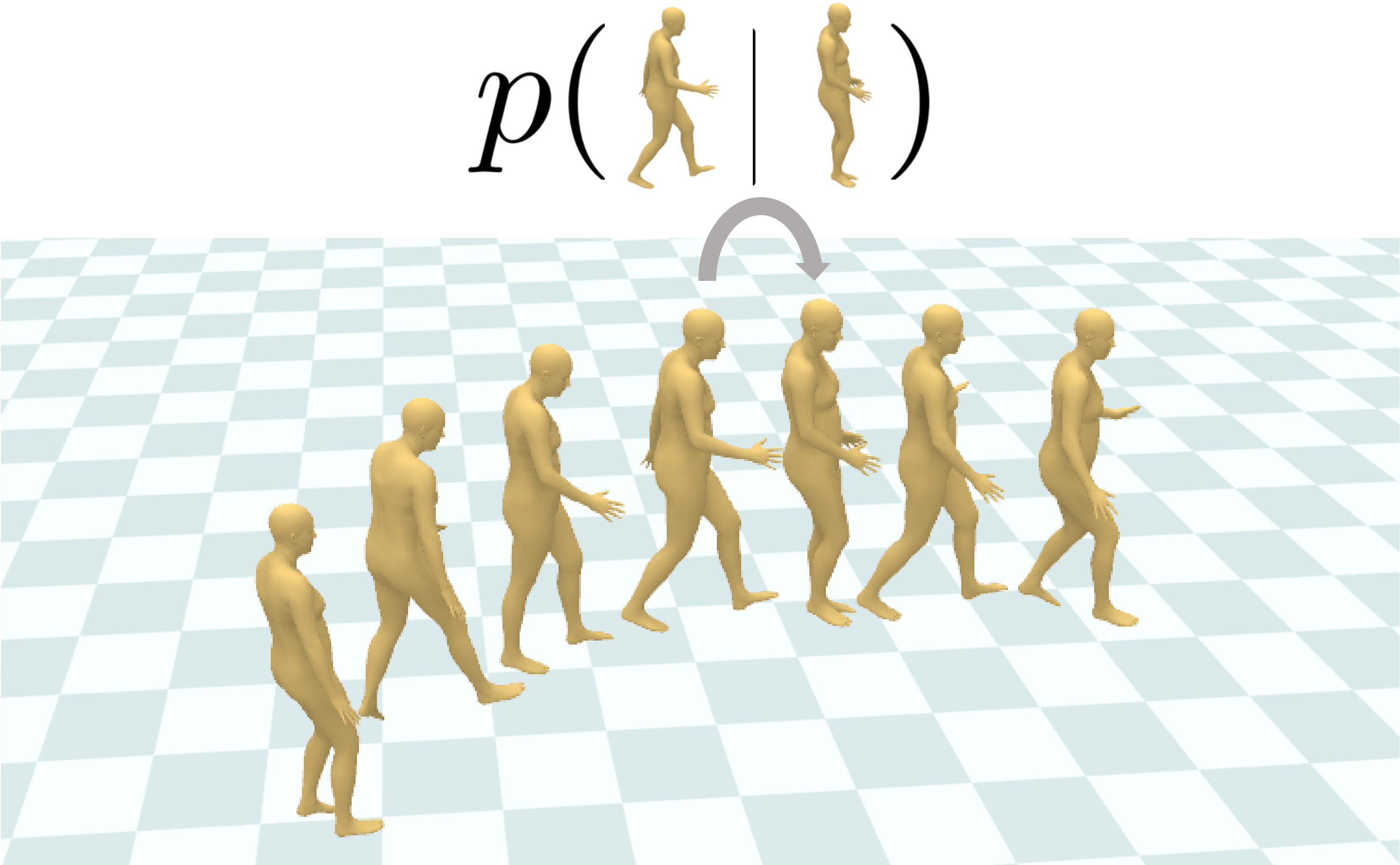


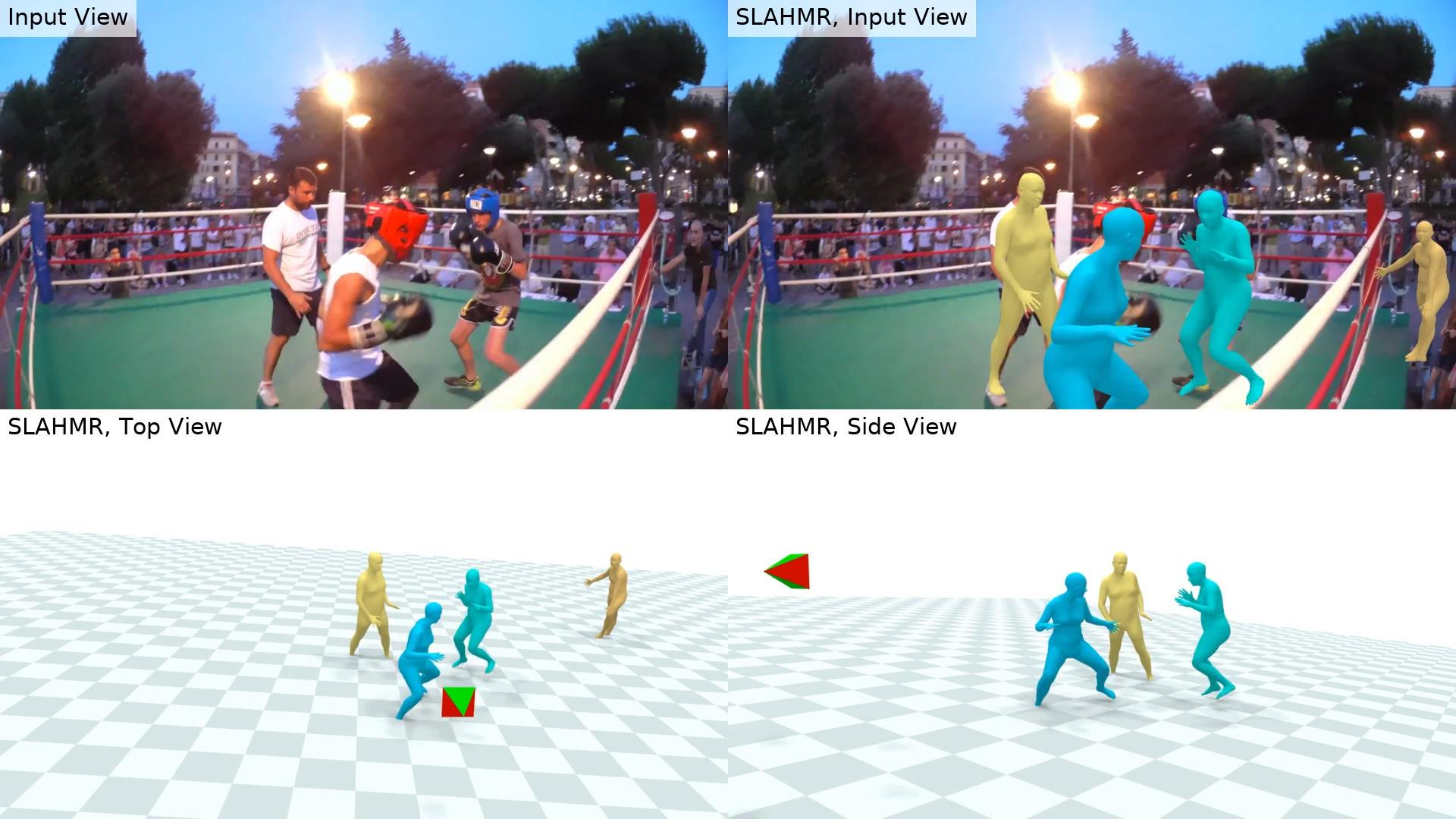
SLAHMR: Simultaneous Localization and Human Mesh Recovery

Key signal: Motion Prior

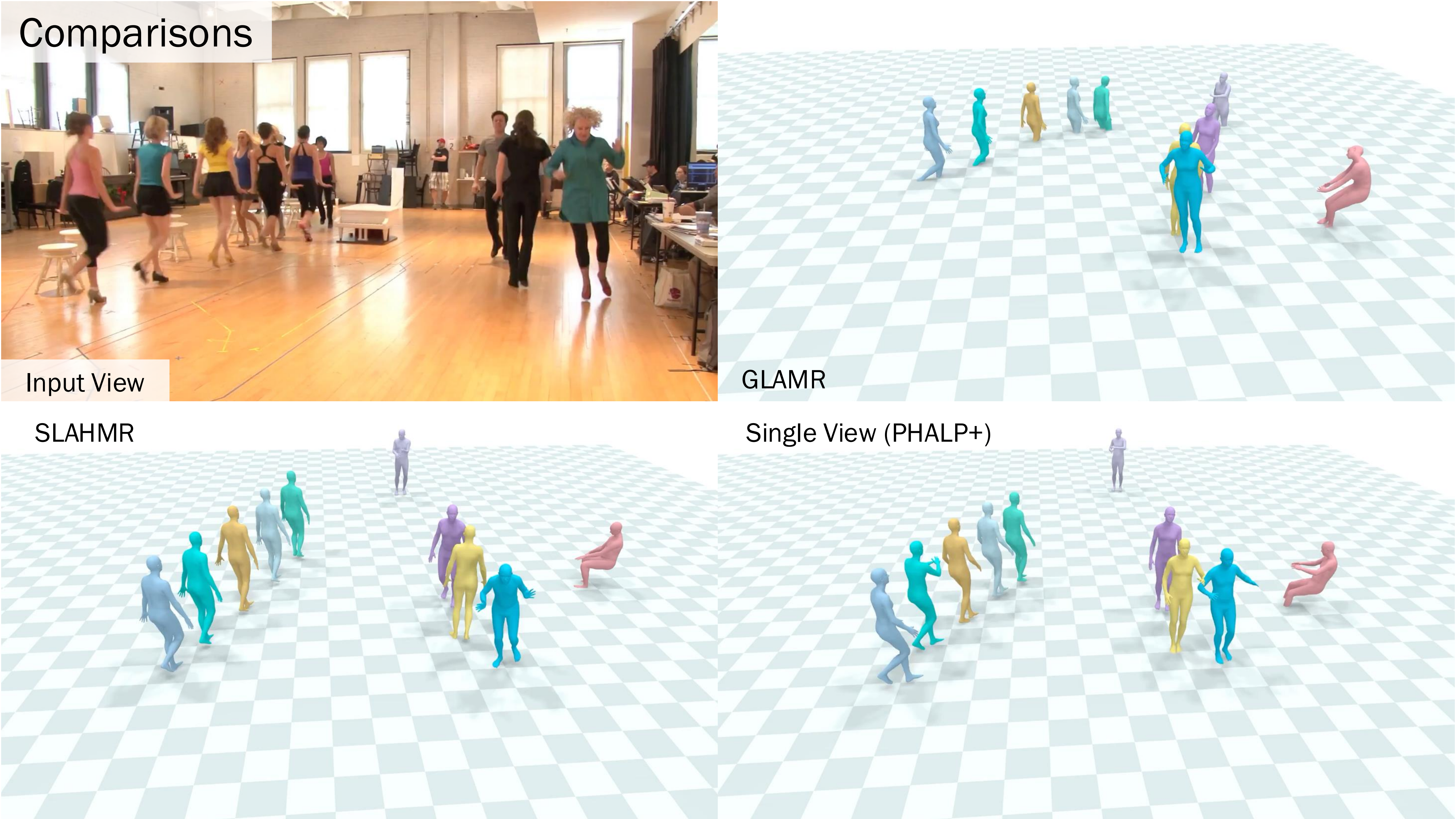
Recover World that gives most probable motion

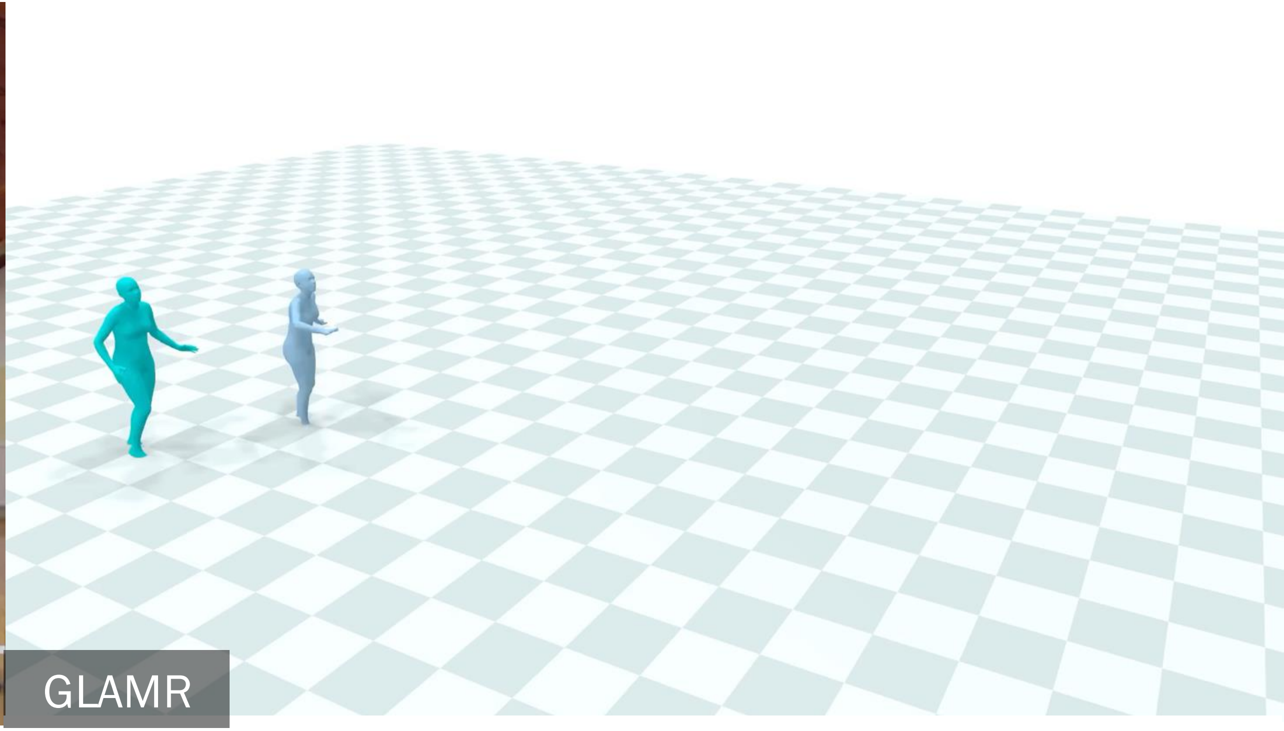
A data-driven motion prior: HuMoR [Rempe et al. ICCV 2021]




Input View
SLAHMR, Input View
SLAHMR, Top View
SLAHMR, Side View

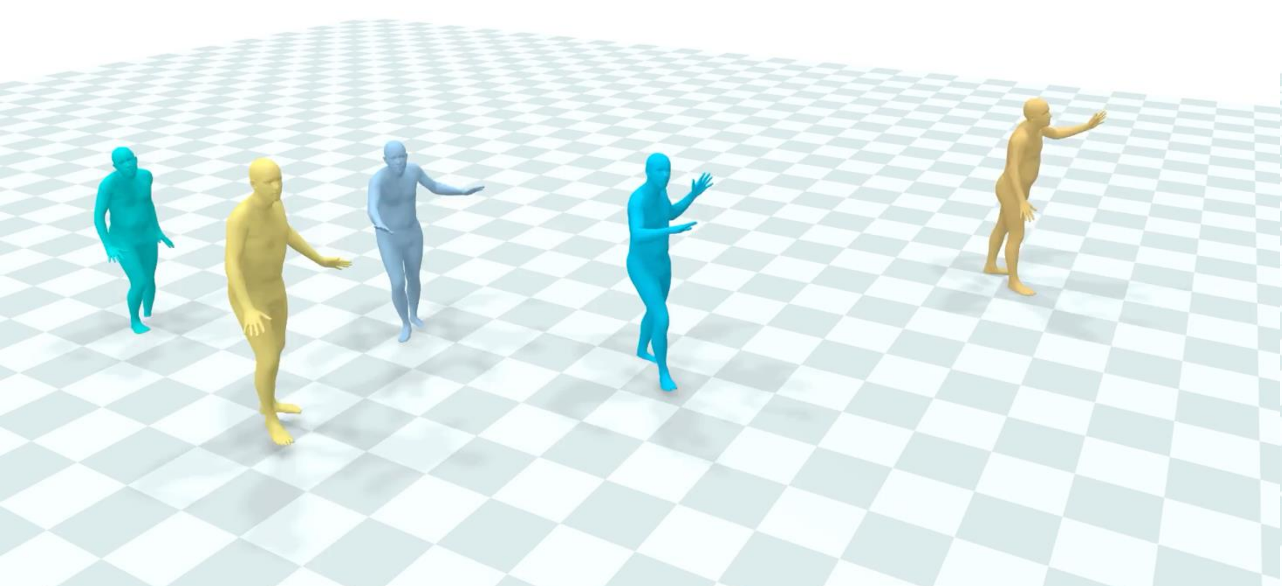

Input View
SLAHMR, Input View
SLAHMR, Top View
SLAHMR, Side View

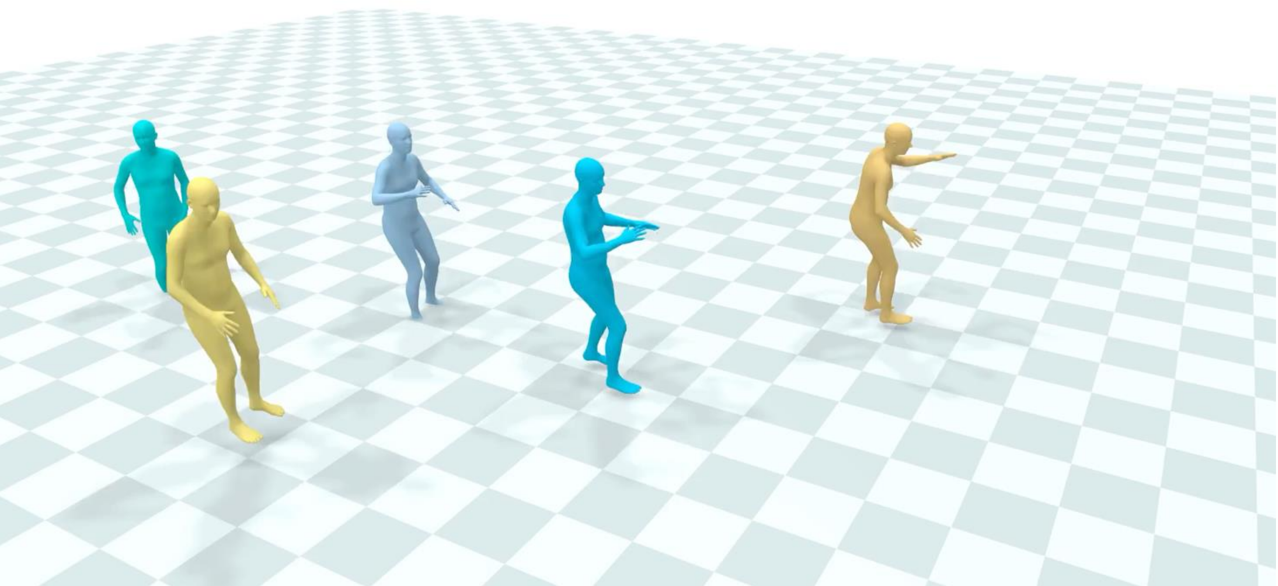

Comparisons
Input View
GLAMR
SLAHMR
Single View (PHALP+)



SLAHMR

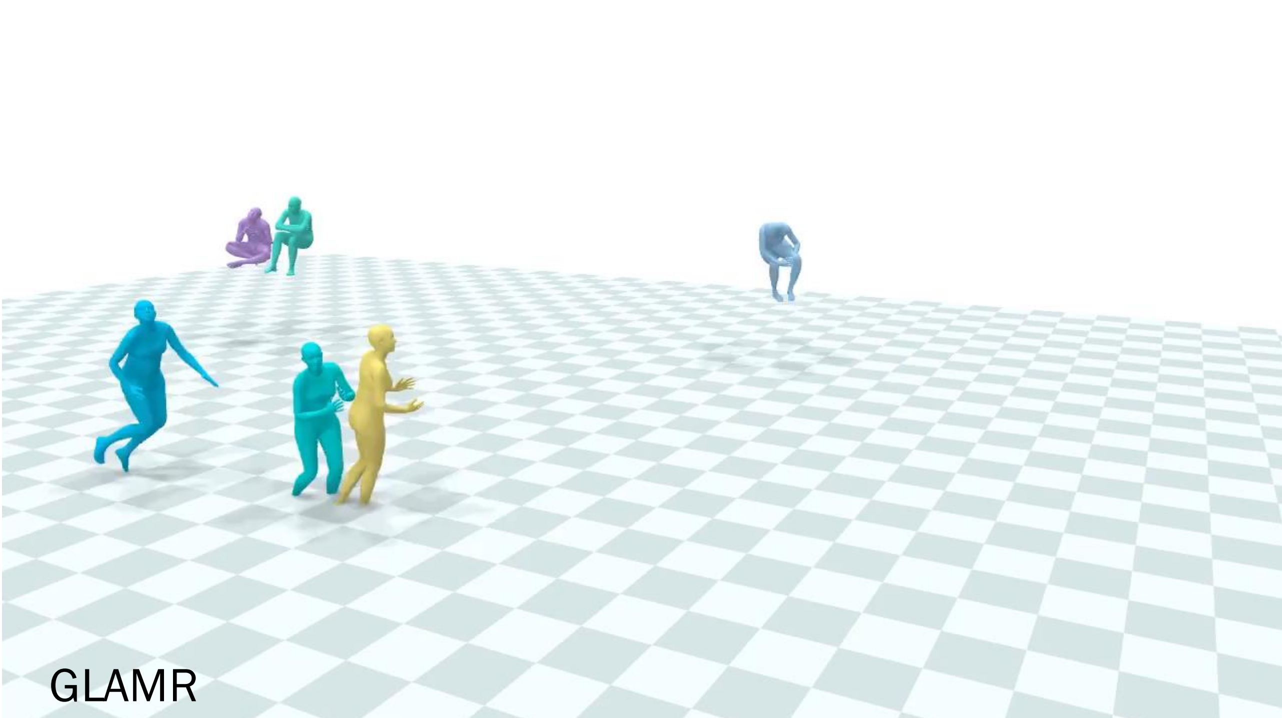
Single View (PHALP+)



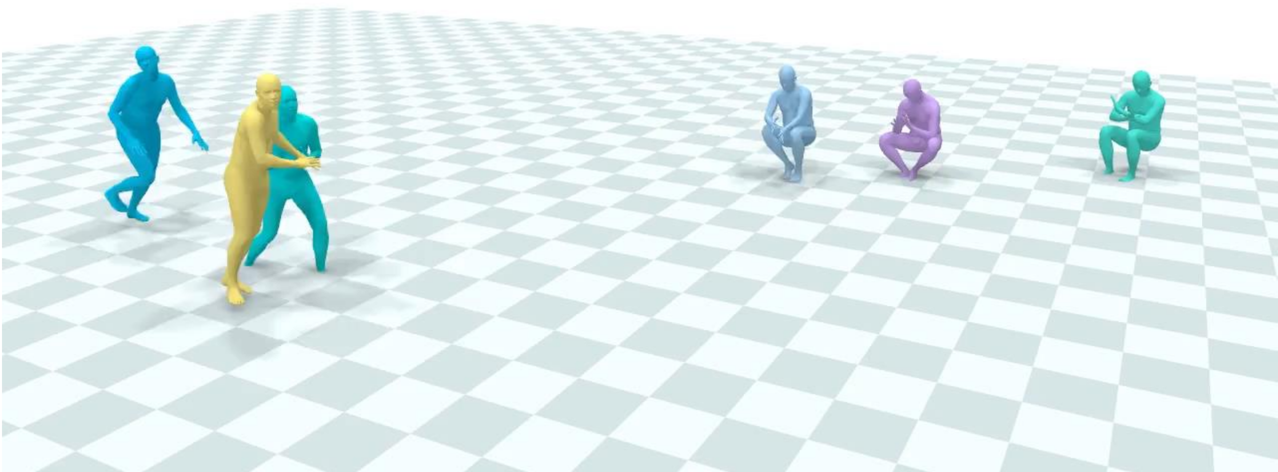


Comparisons

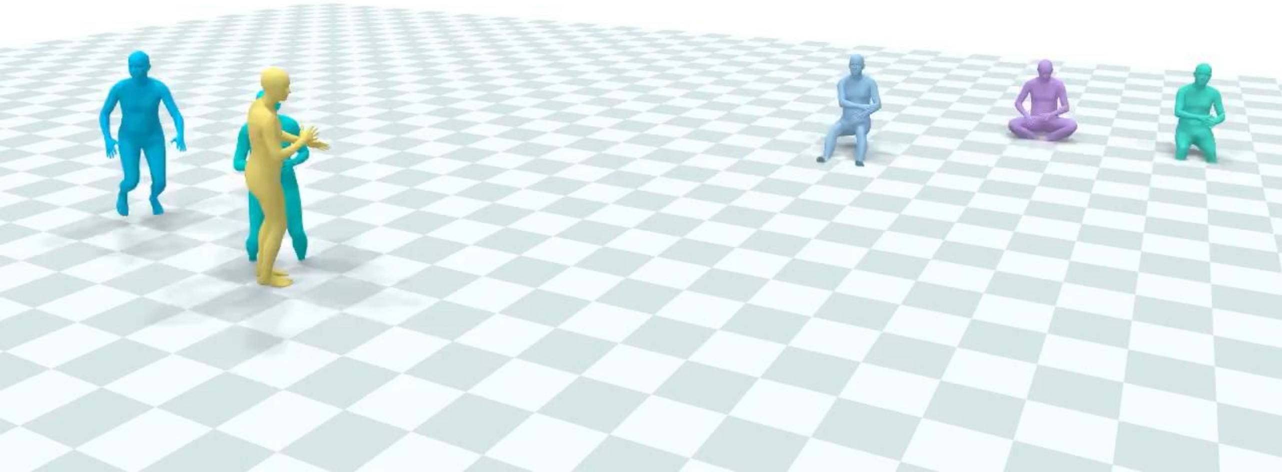

Input View


GLAMR

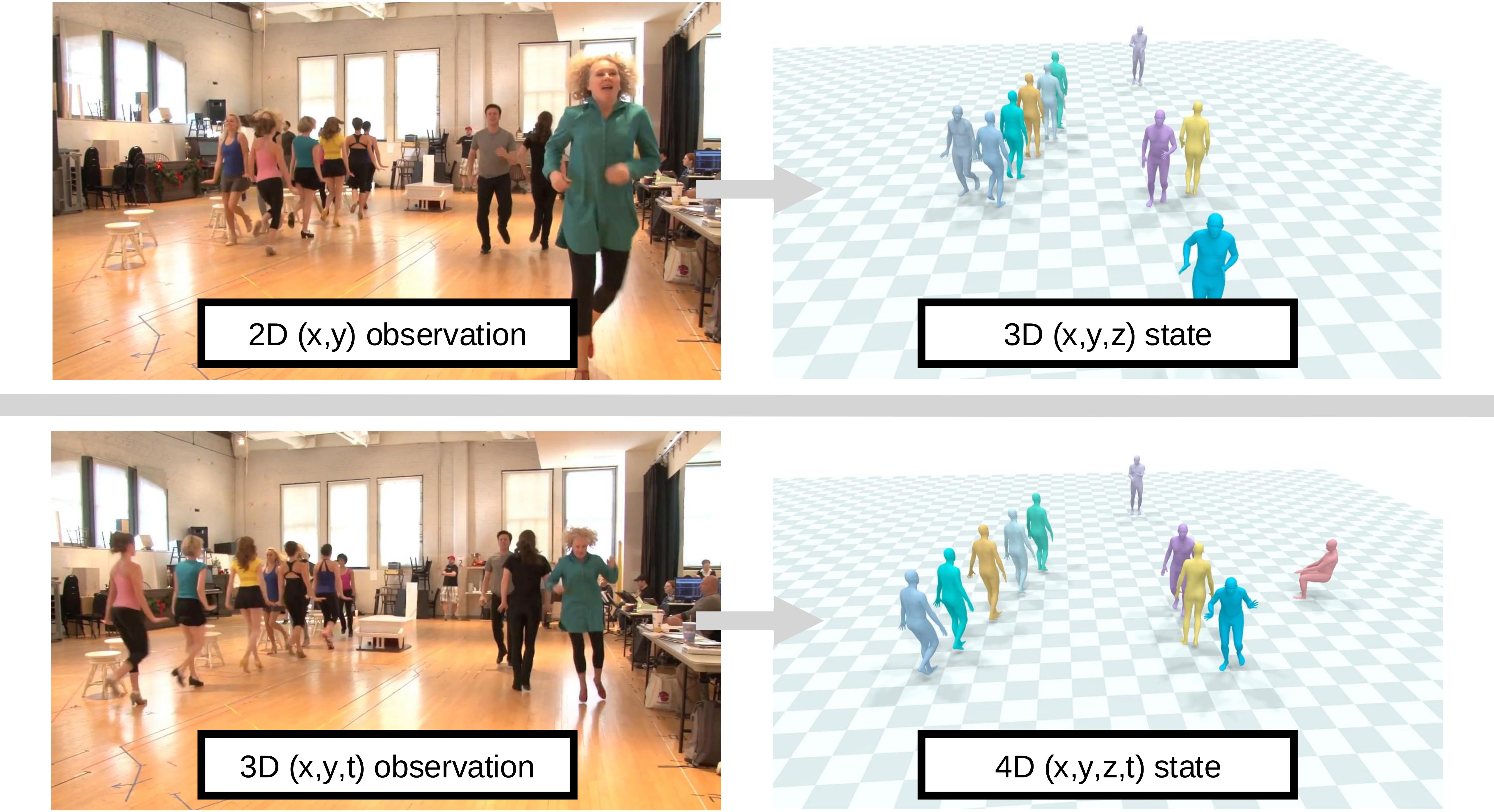
Ours



Single View (PHALP+)



4D Reconstruction


2D (x,y) observation
3D (x,y,z) state
3D (x,y,t) observation
4D (x,y,z,t) state

Prereq: tracking people across time

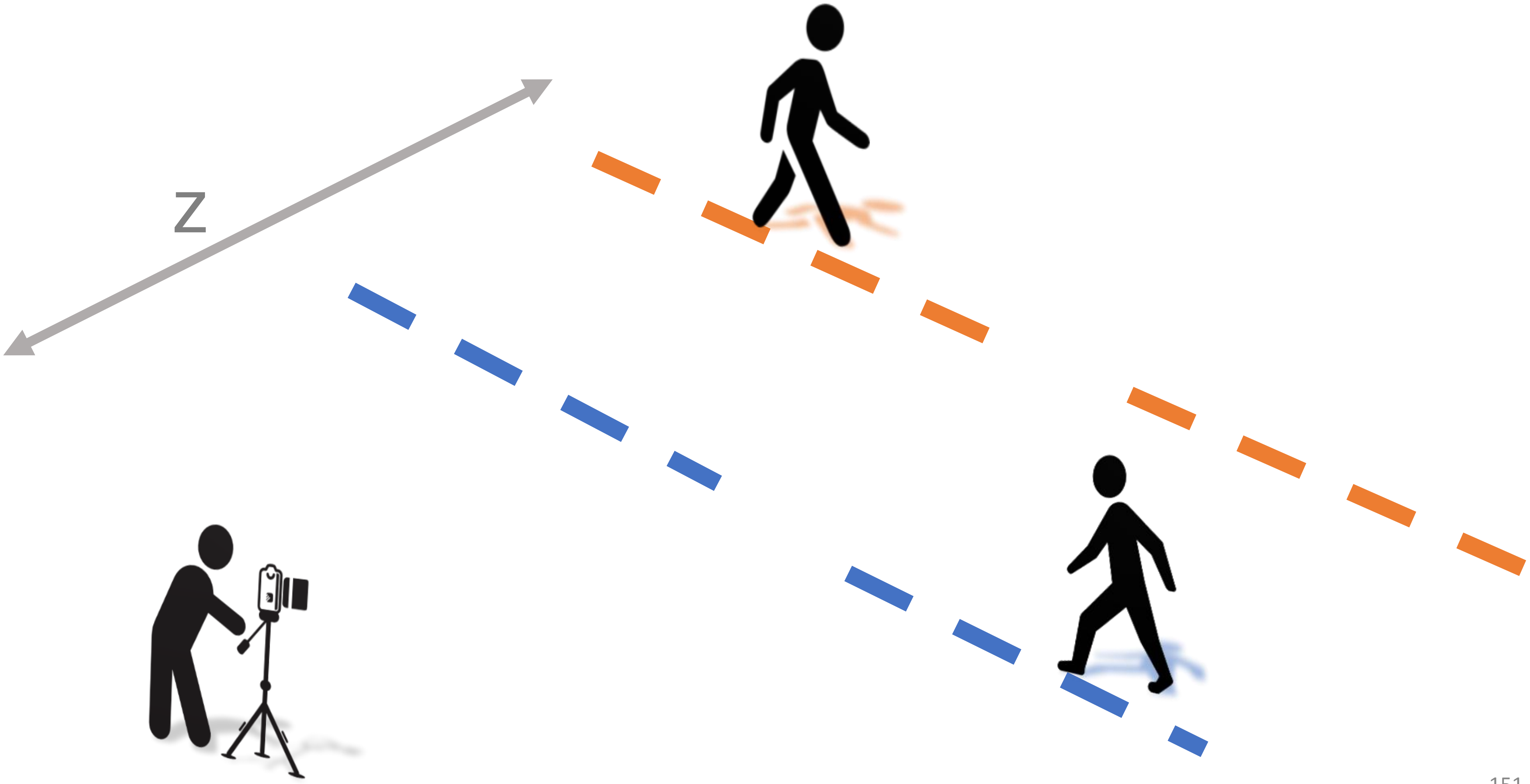
PHALP: Rajasegaran, Pavlakos, Kanazawa, Malik, CVPR 2022 (Oral)

- In 2D, occlusion is hard to disambiguate

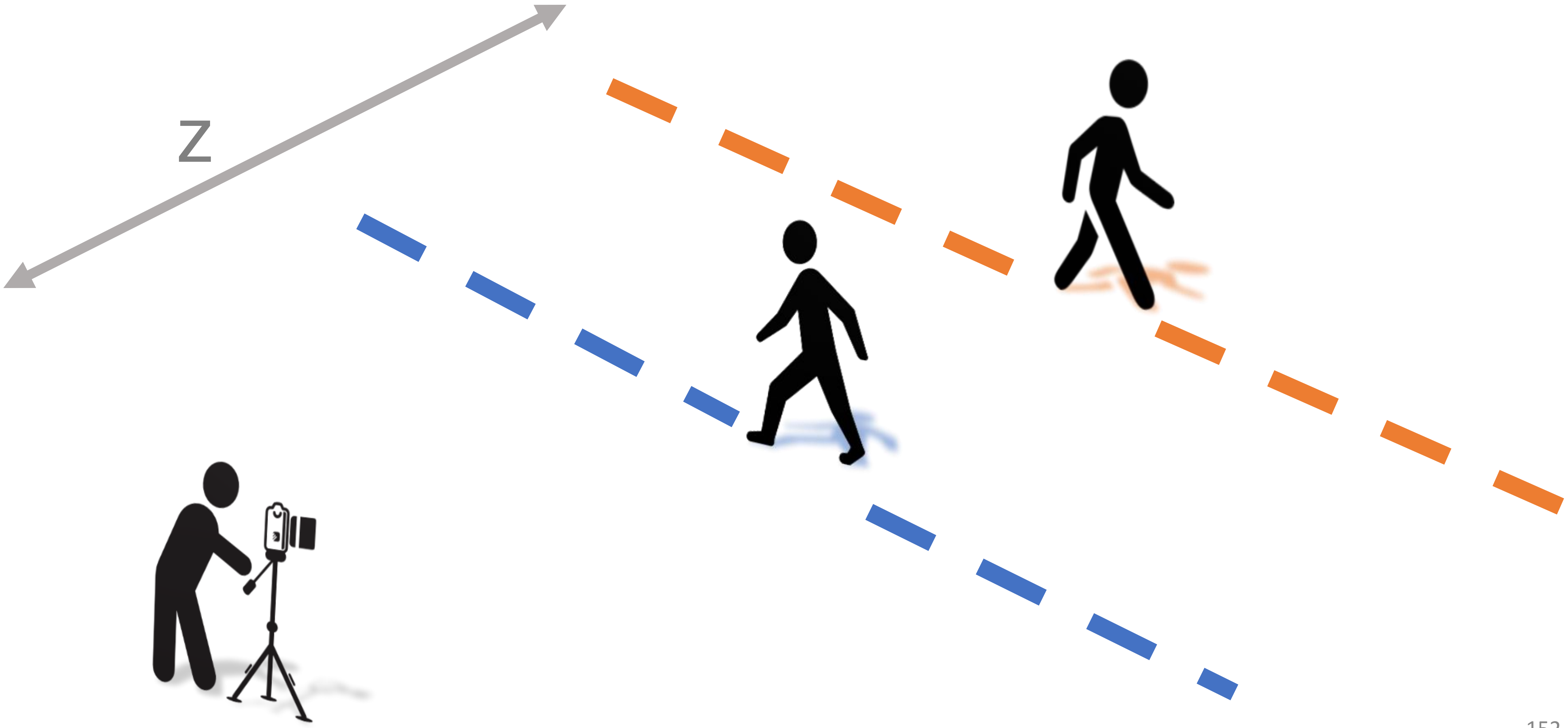
- In 2D, occlusion is hard to disambiguate

- In 2D, occlusion is hard to disambiguate

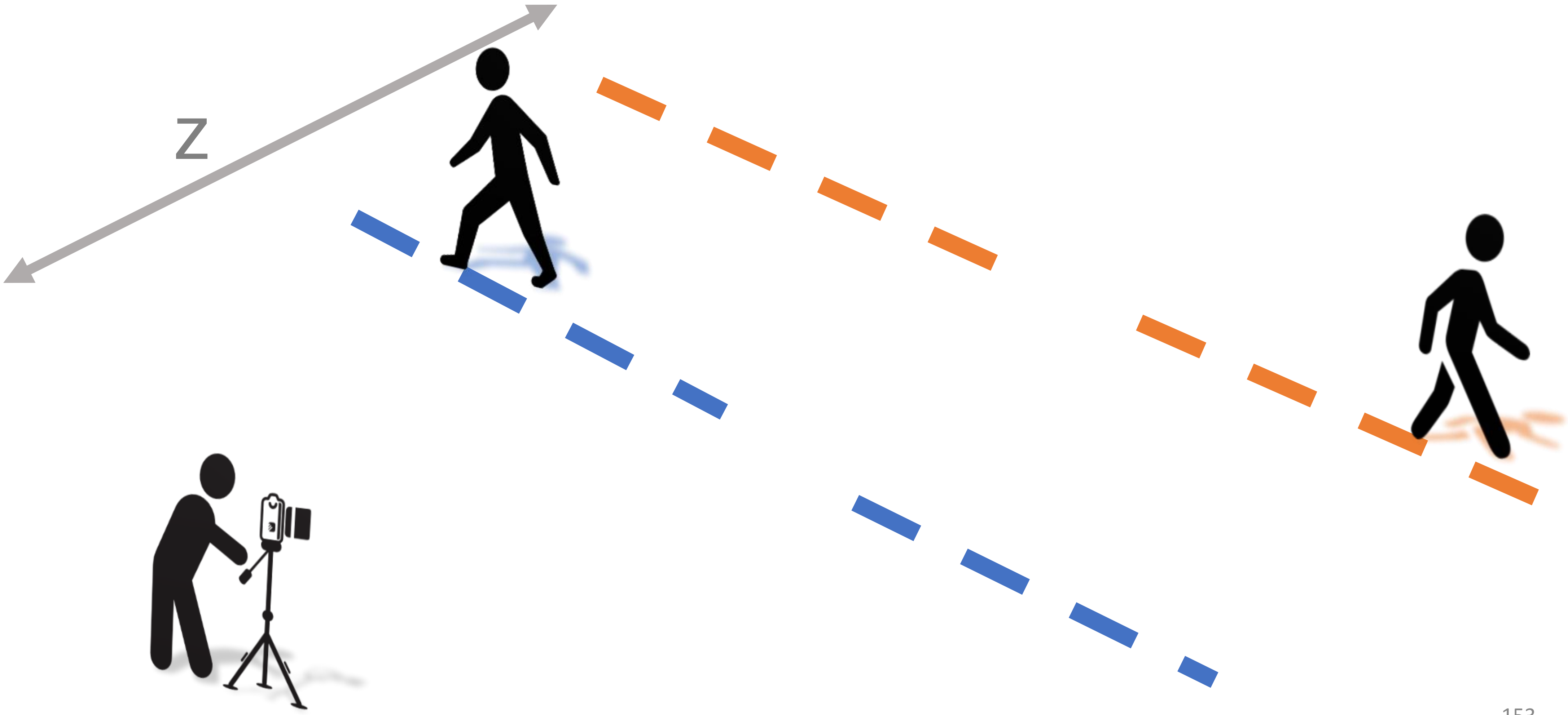
- In 2D, they overlap, but in 3D they don't!

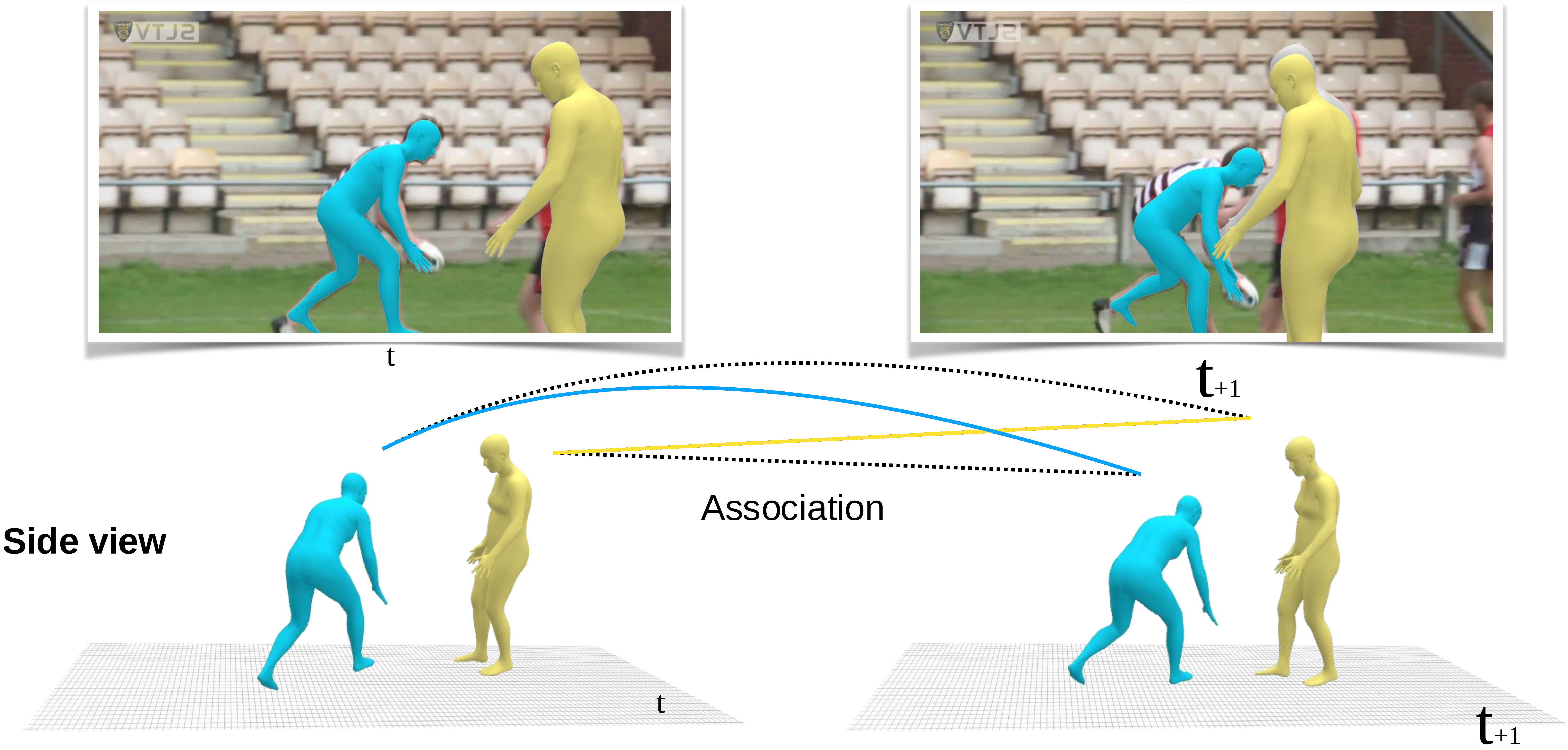


- In 2D, they overlap, but in 3D they don’t!



- In 2D, they overlap, but in 3D they don't!





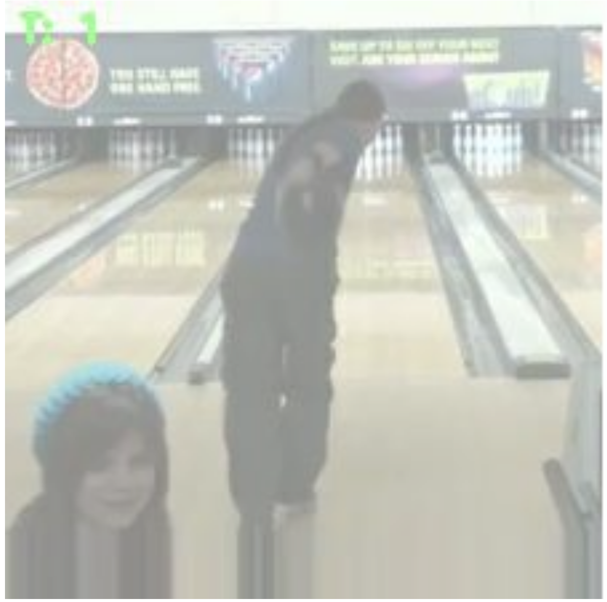
Distance = 3D location distance + 3D appearance distance + 3D pose distance

PHALP: Rajasegaran, Pavlakos, Kanazawa, Malik, CVPR 2022 (Oral)

A benefit of video: Dynamics

Auto-regressive prediction of 3D motion from video

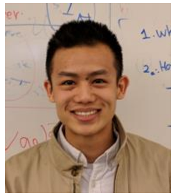
Input Video



Ground Truth Video

Predicted Future

Different Viewpoint



Predicting 3D Human Dynamics from Video, Zhang, Felsen, Kanazawa, Malik, ICCV 2019

Test Time

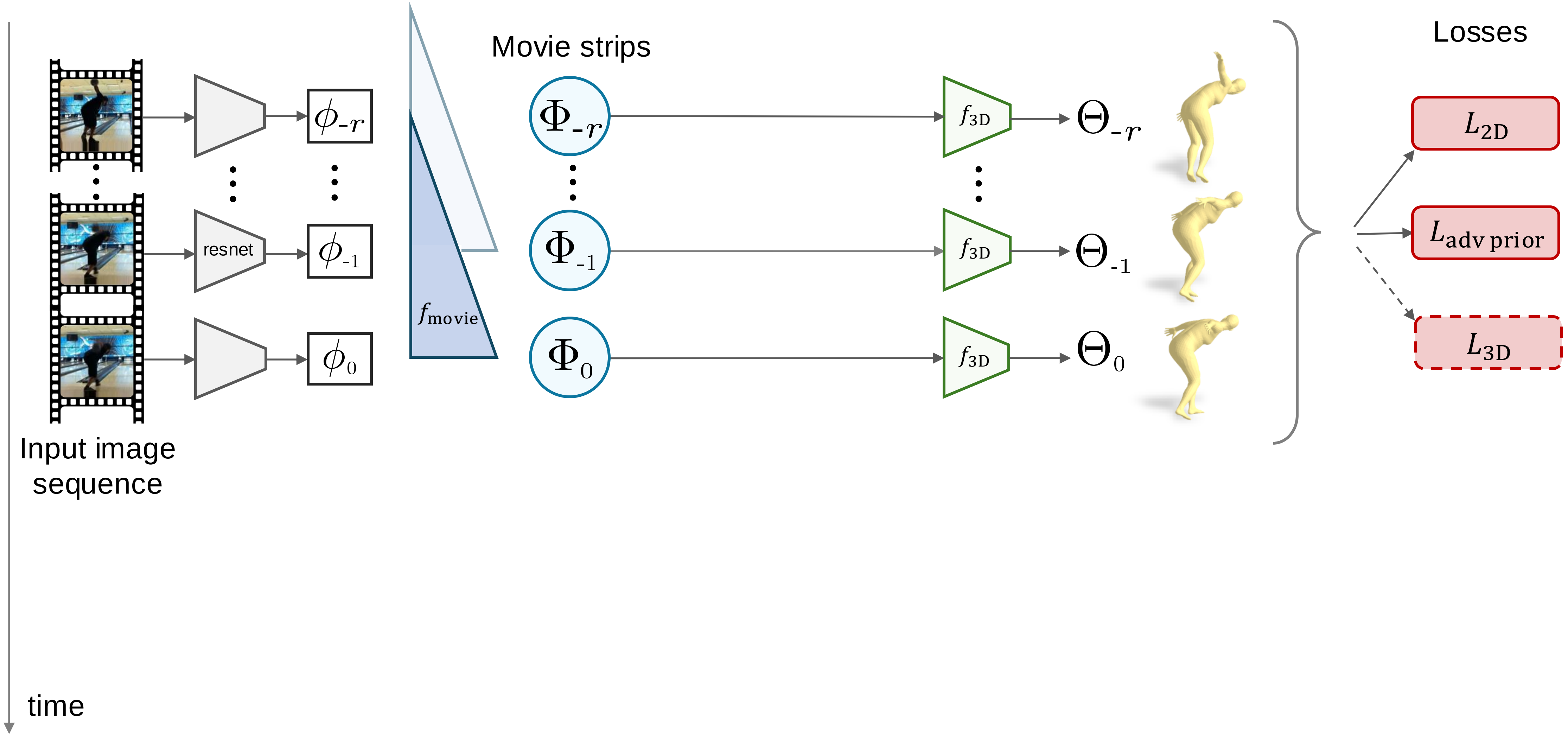
Input Video

Ground Truth Video

Predicted Future

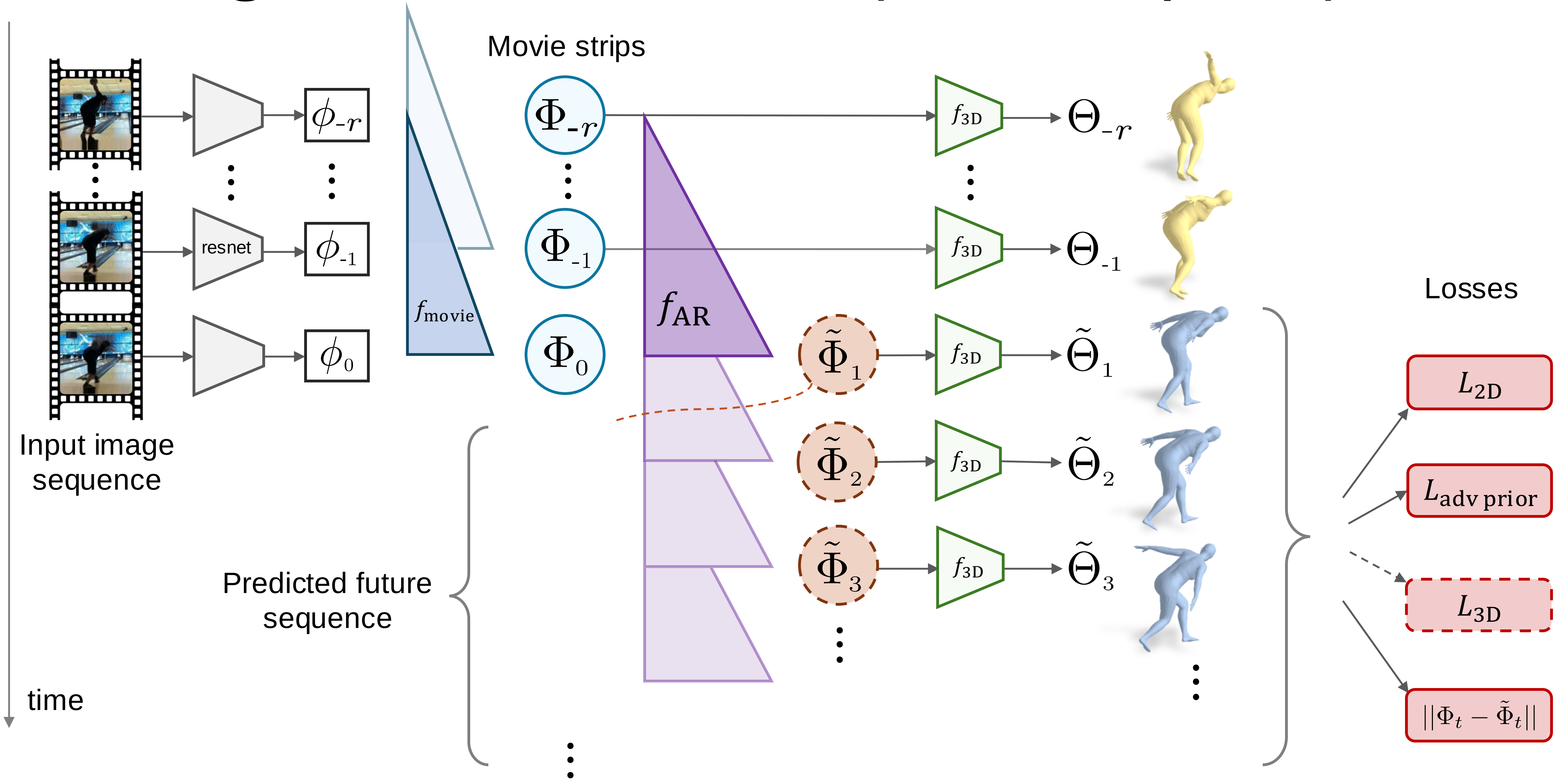
Different Viewpoint

Overview



Movie strips
Losses
resnet
ϕ_{-r}
ϕ_{-1}
ϕ_0
f_{movie}
Φ_{-r}
Φ_{-1}
Φ_0
f_{3D}
Θ_{-r}
Θ_{-1}
Θ_0
L_{2D}
$L_{\text{adv prior}}$
L_{3D}
Input image sequence
time

Autoregressive Prediction (Latent Space)



Yellow = Conditioning

Blue = Future Prediction from movie strip

Camera View

Alternate Viewpoint

Yellow = Conditioning

Blue = Future Prediction from movie strip

Camera View

Alternate Viewpoint

Yellow = Conditioning

Blue = Future Prediction from movie strip

Camera View

Alternate Viewpoint

Finally, a step towards this baby

SfV: Reinforcement Learning of Skills from Videos

Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, Sergey Levine

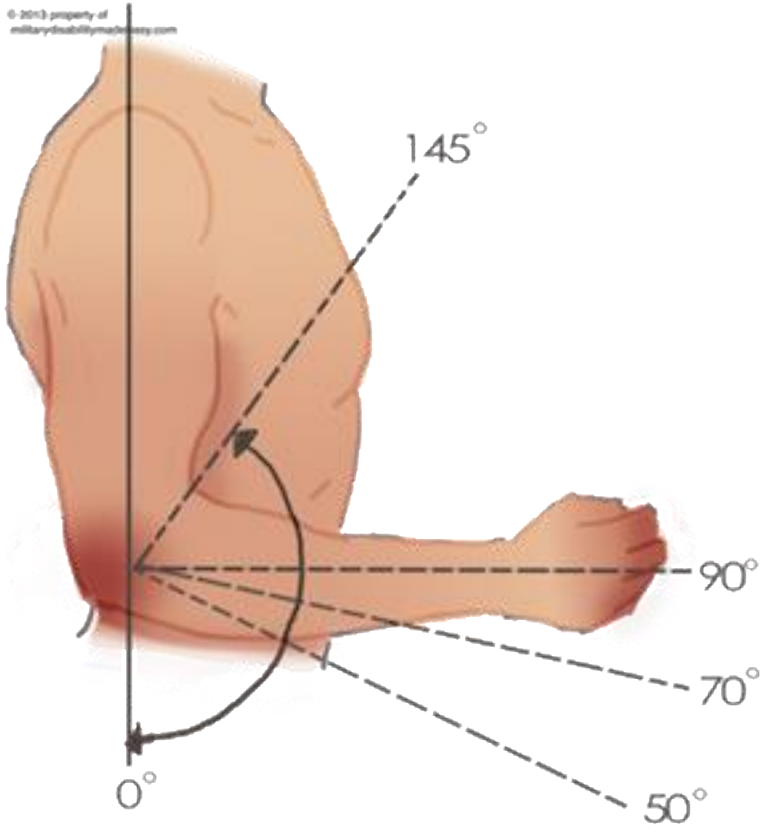
SIGGRAPH Asia 2018



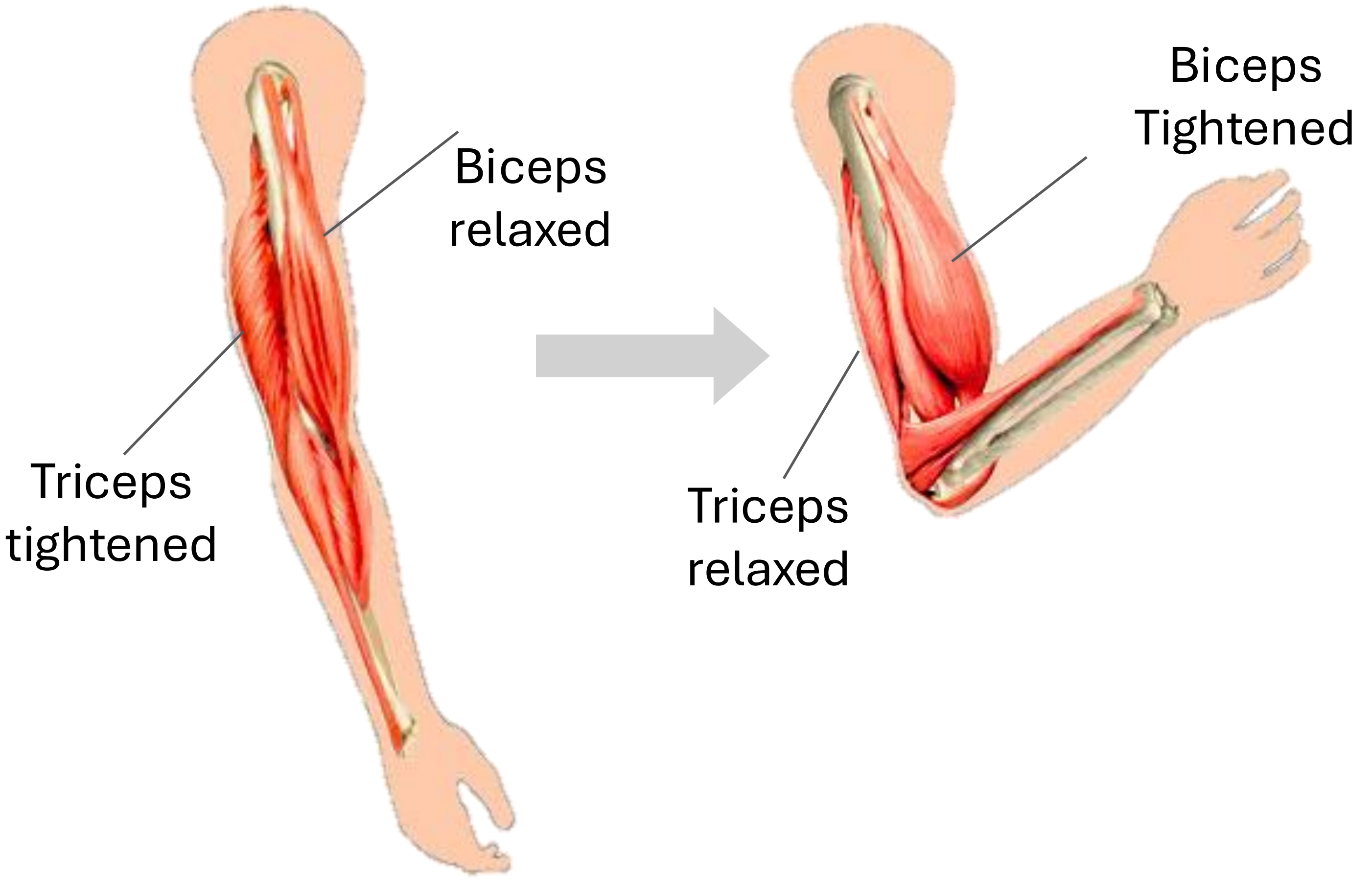


Perception is not the end of the story

Perceiving the 3D pose

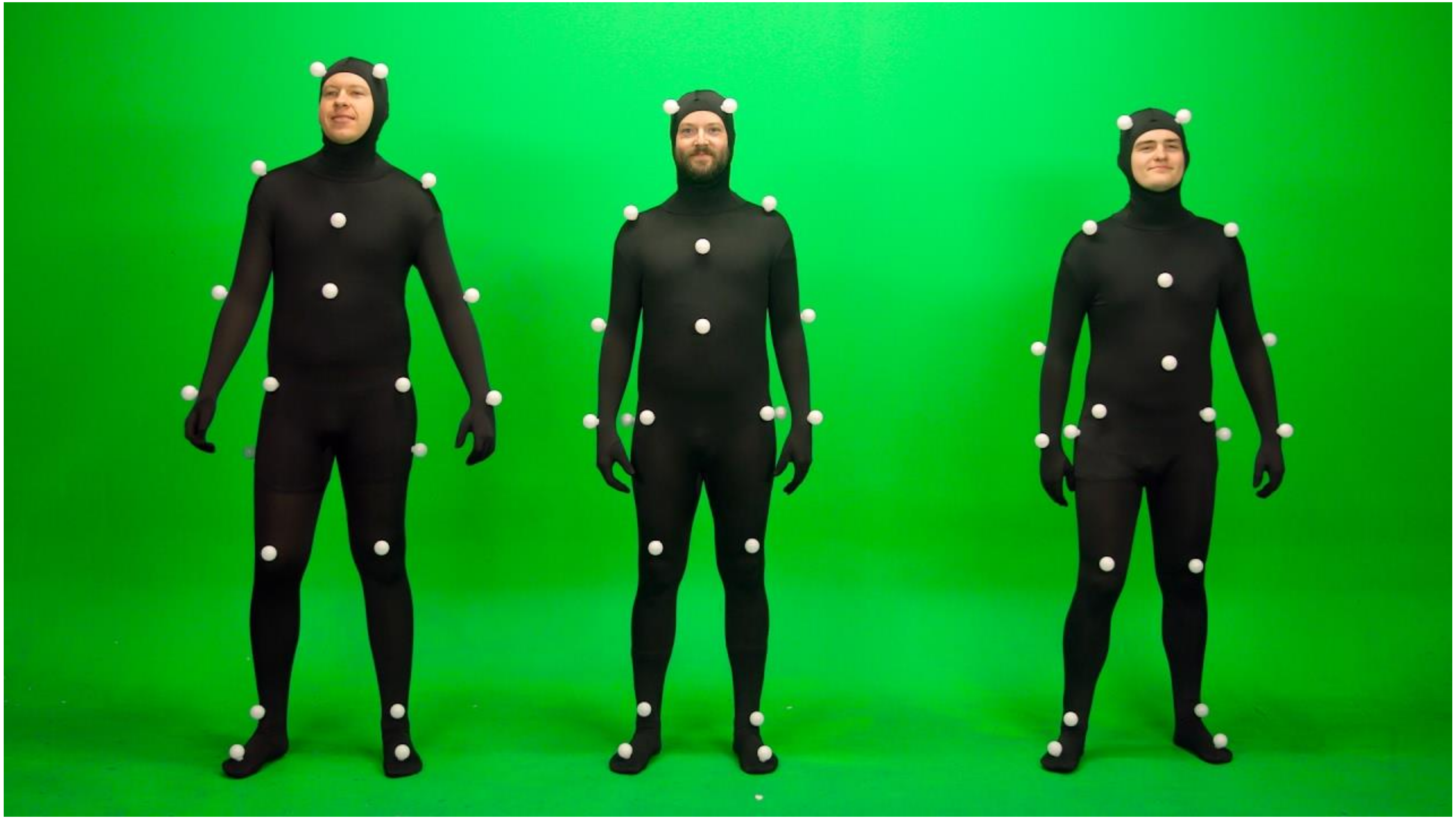


Actuating the muscle to get to the pose

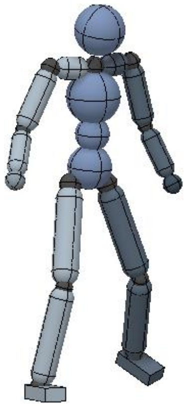


Control is not easy!

Past work on start from mocap

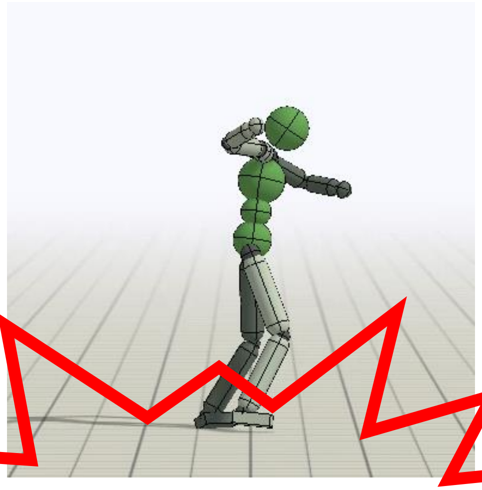


Deep Reinforcement Learning Based Motion Imitation



Character
In simulation

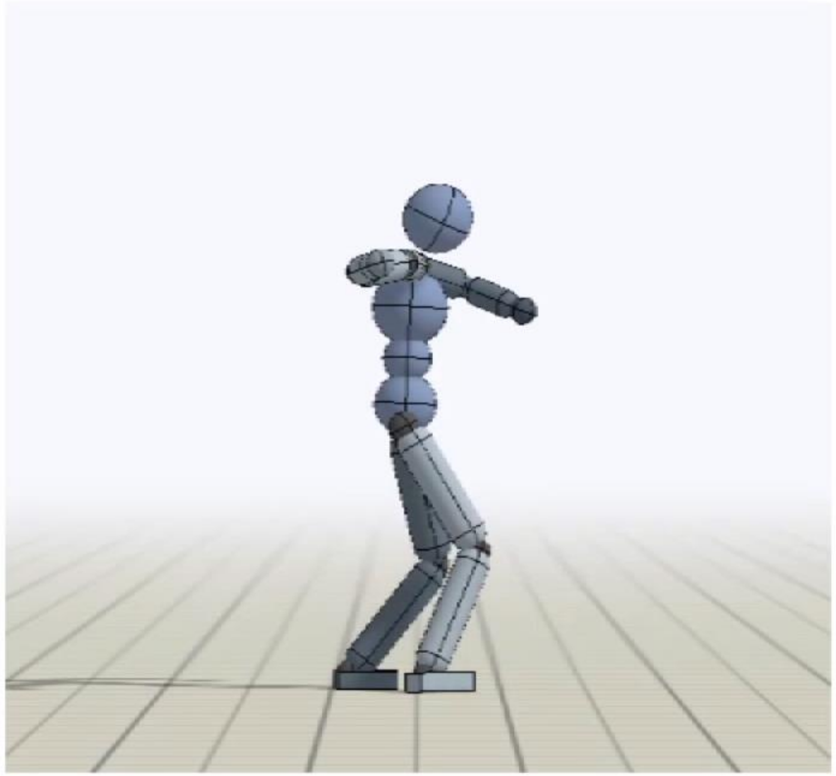
Need Actions



Reference Motion
(**Mocap**)

Only a
recording of
joint angles

Reinforcement
Learning



DeepMimic [Peng et al. SIGGRAPH 2018]

Expand the world to video

YouTube

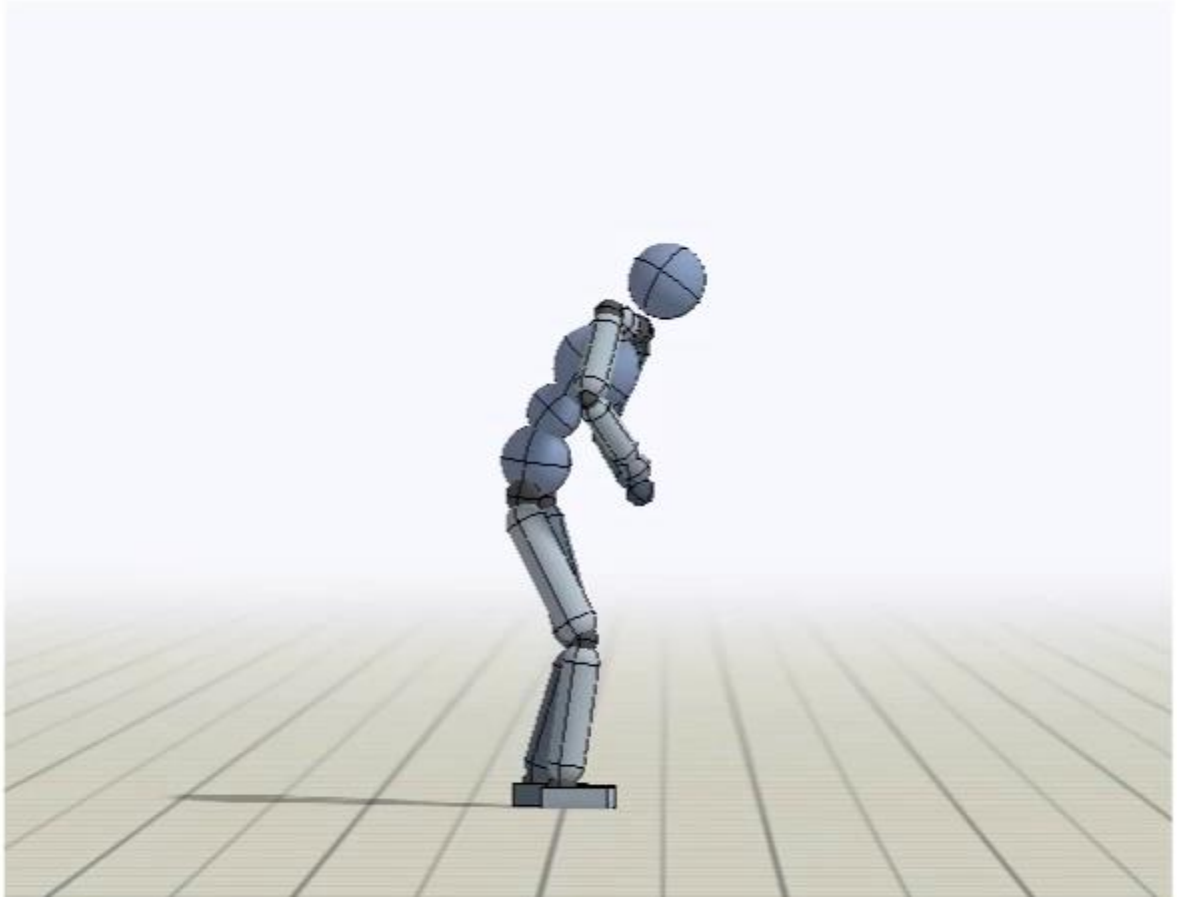
Learning Dynamic Skills from Videos

Video

Temporally smooth mesh recovery

Use recovered 3D pose to train a physically simulated agent

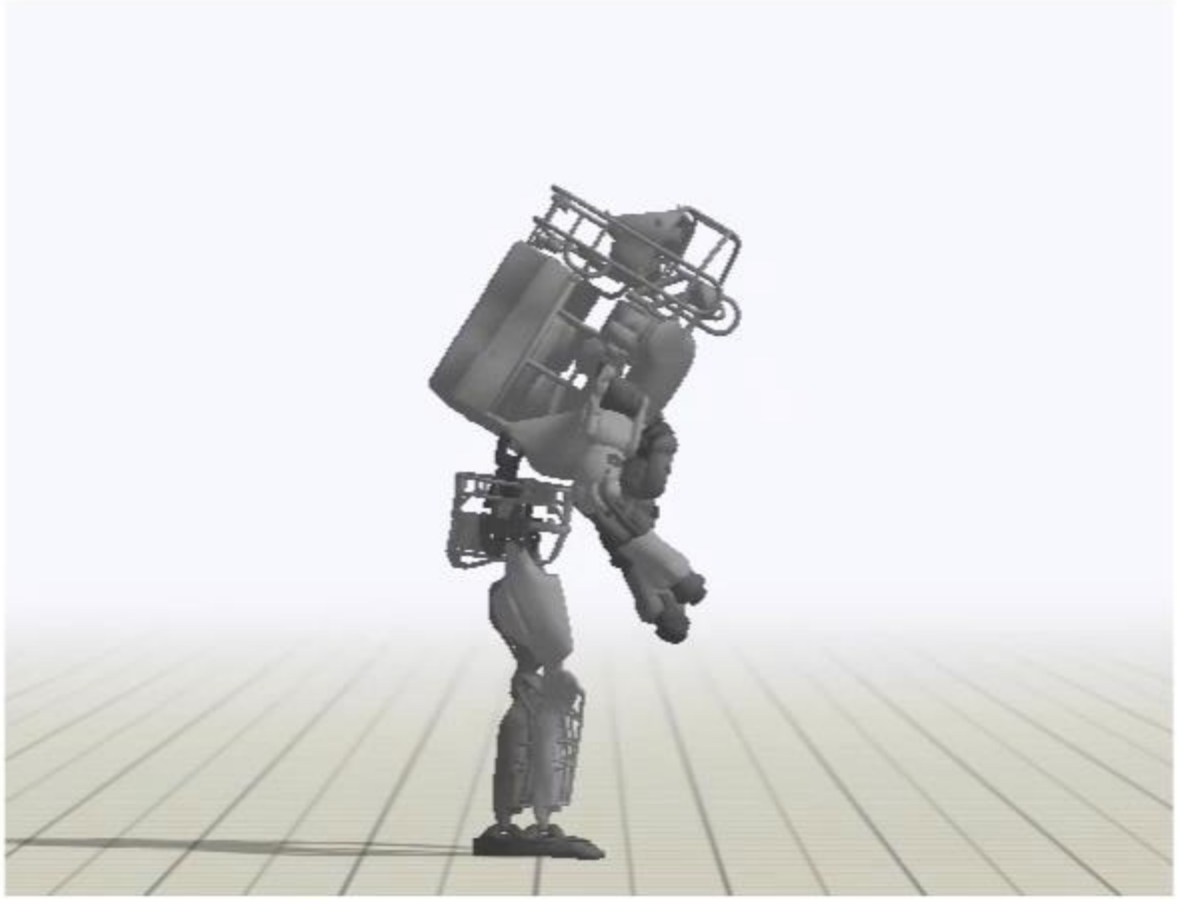
Video



Policy

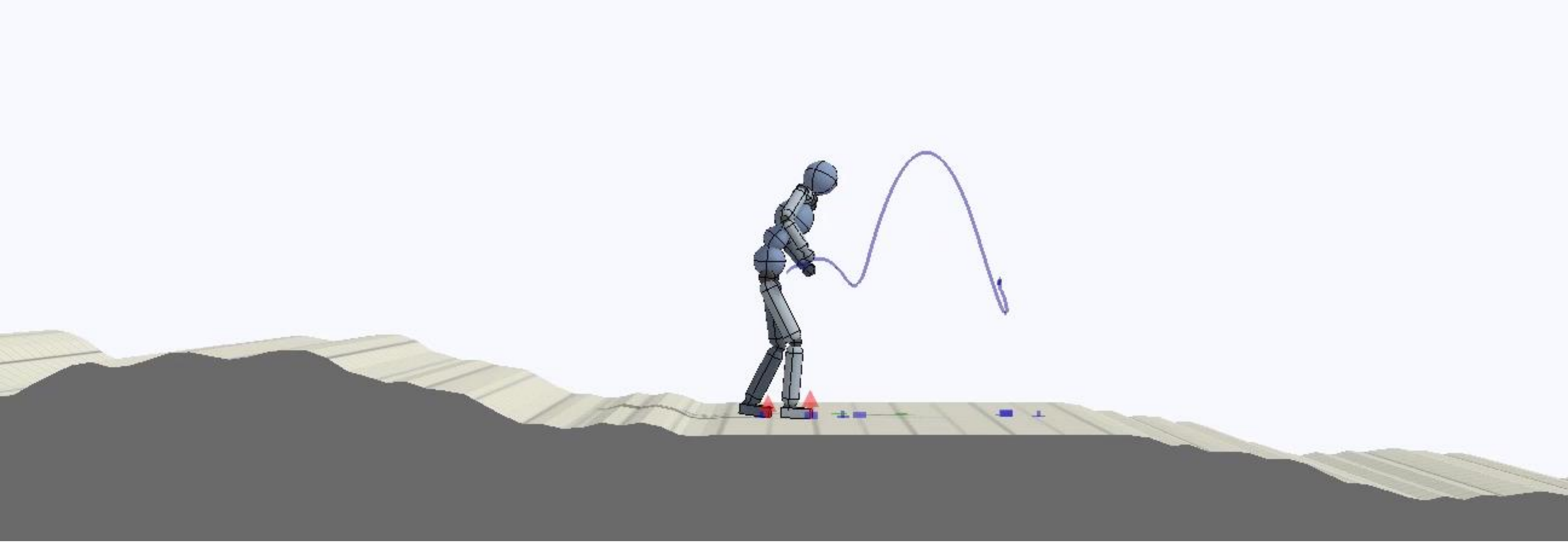
Train Atlas (169kg)

Video

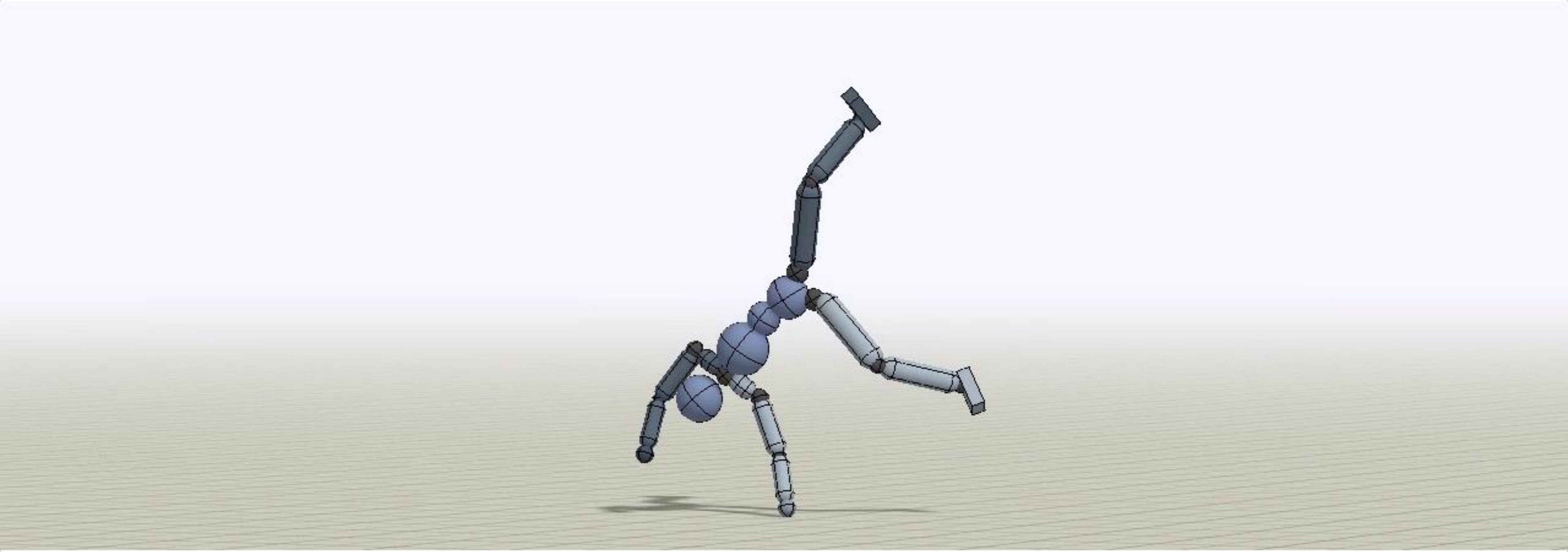


Policy

Environment Retargeting

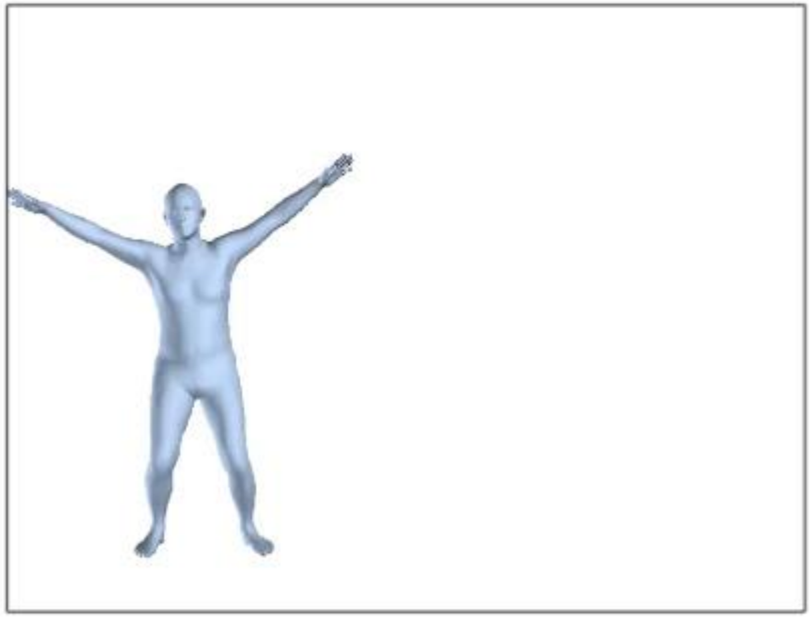


Robustness



Humanoid: Cartwheel

Video

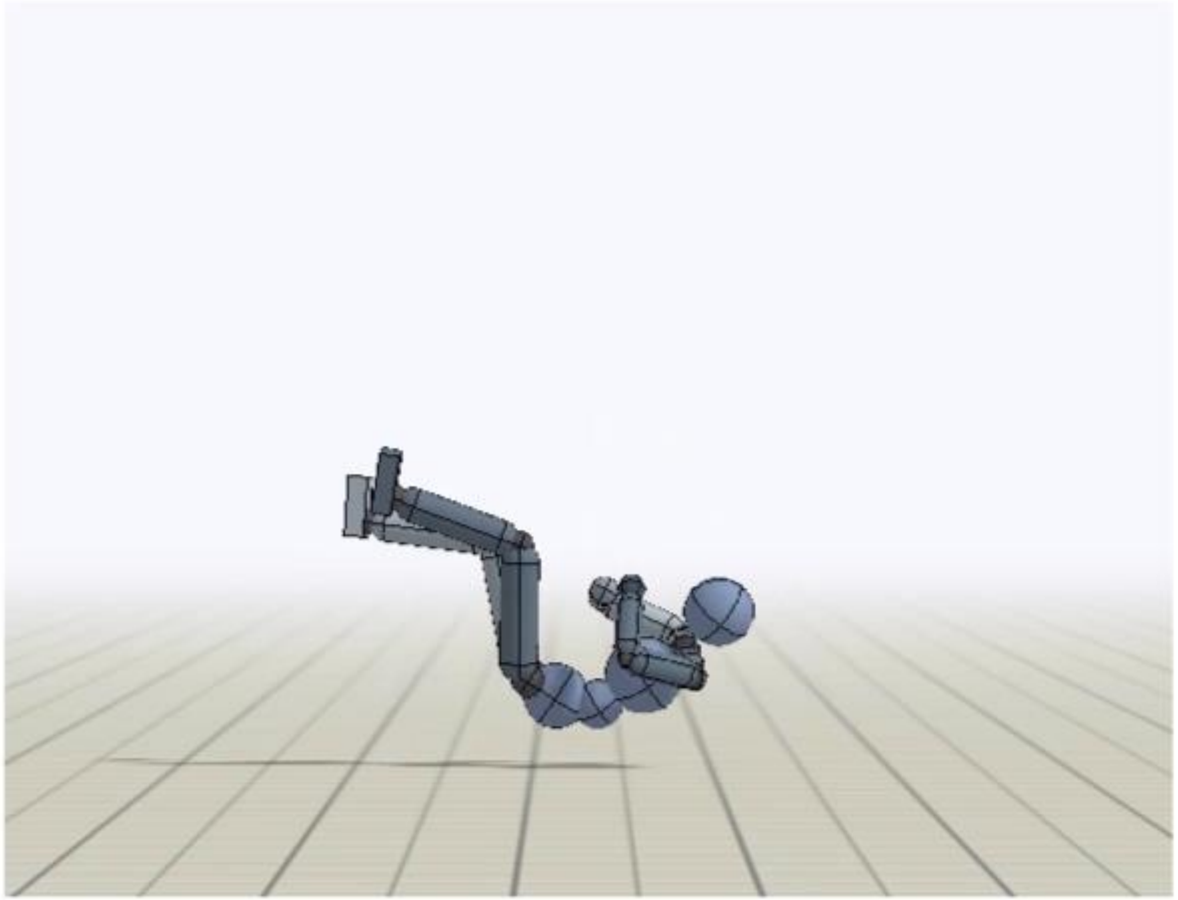


Recovered 3D Body

Policy

Humanoid: Kip-Up

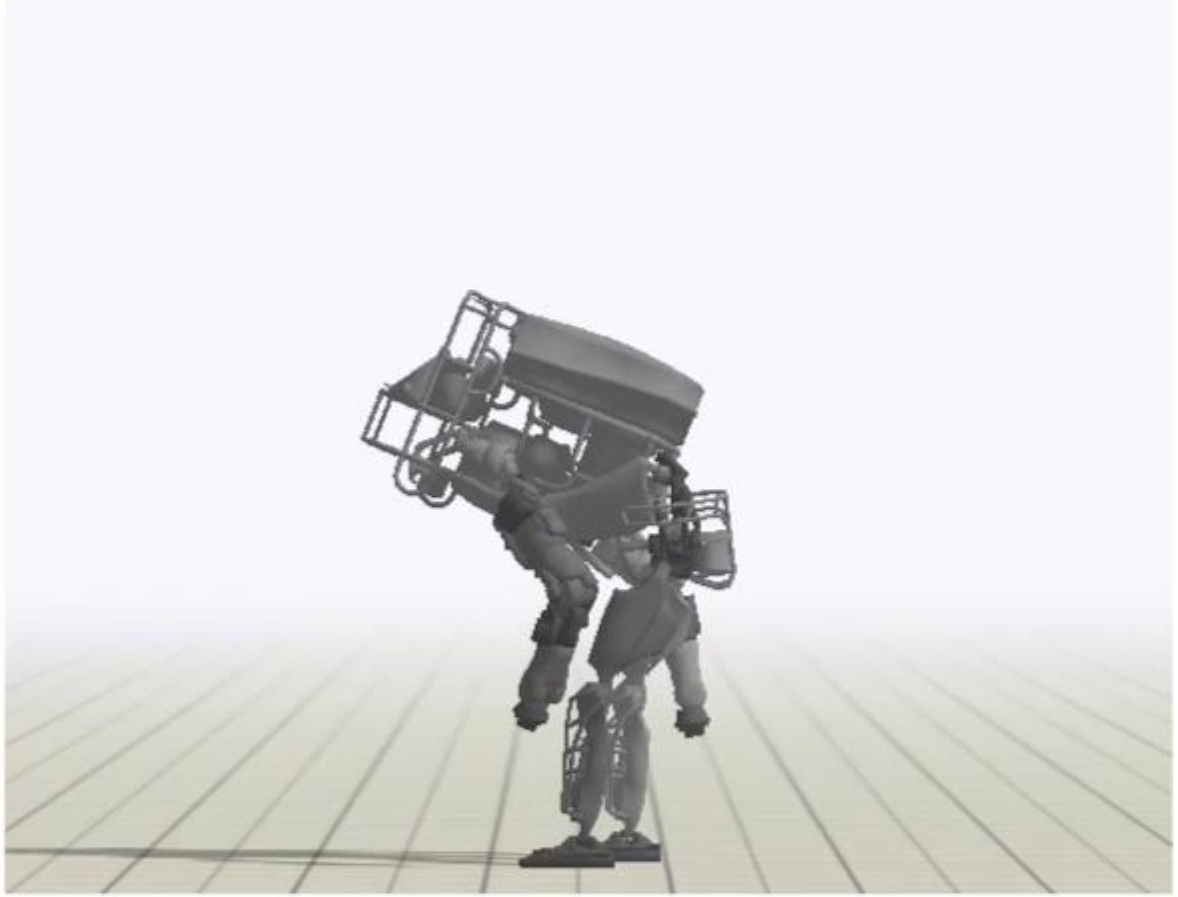
Video: Kip-Up



Policy

Atlas (169kg) : Handspring

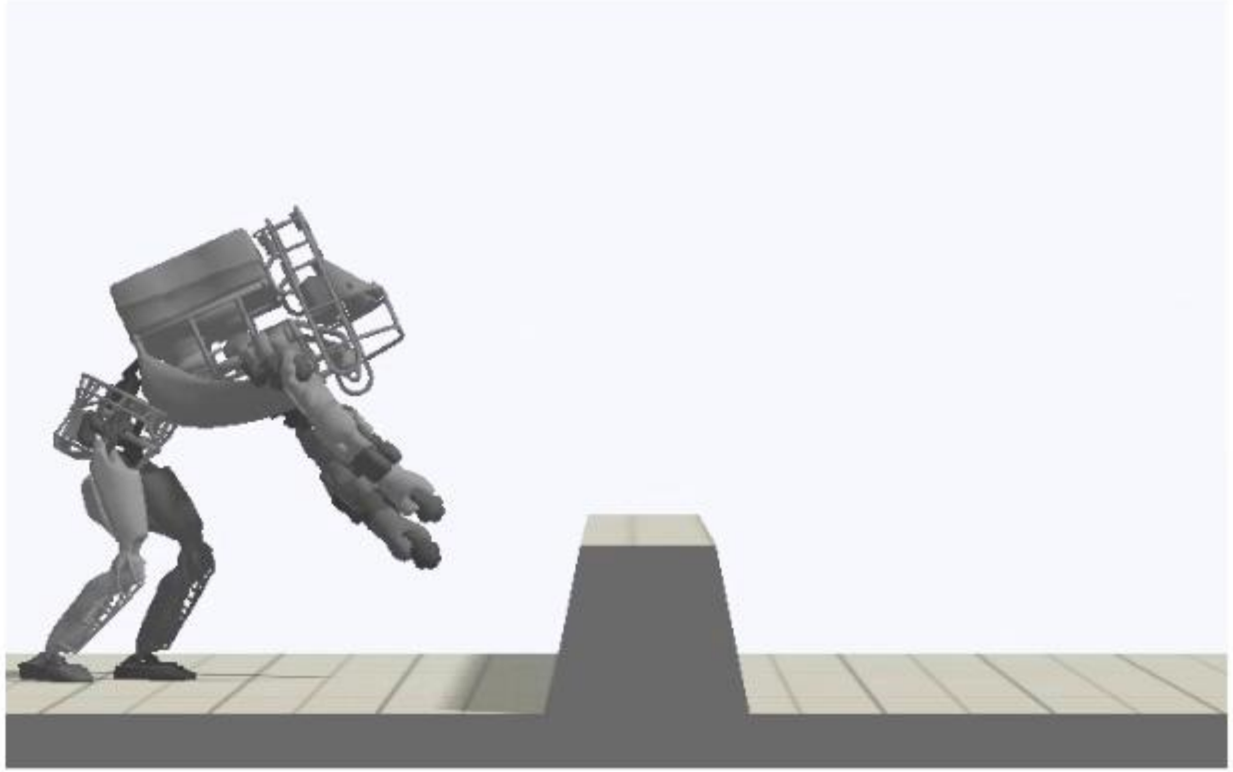
Video: Handspring A



Policy

Atlas: Vault

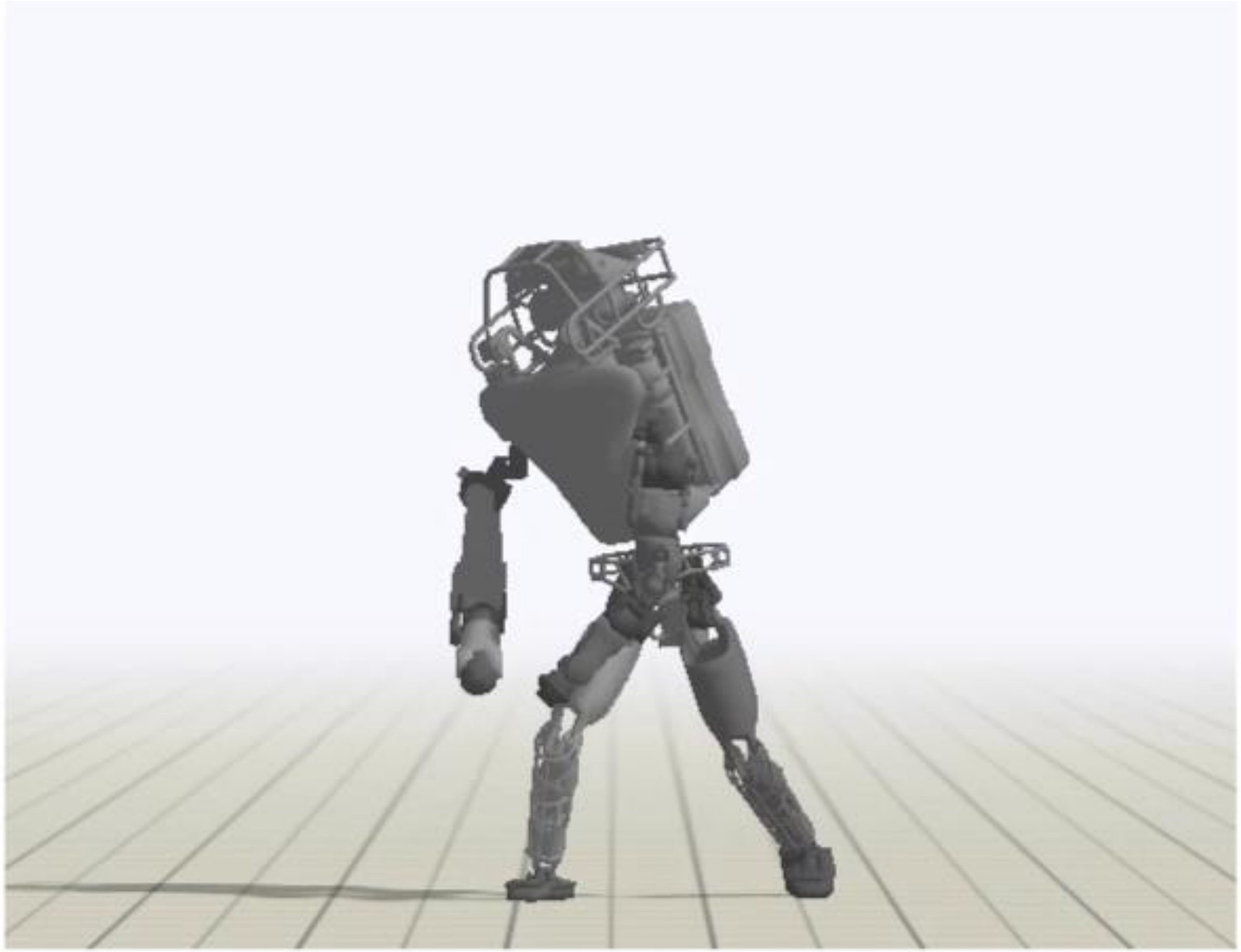
Video: Vault



Policy

Atlas: Dance

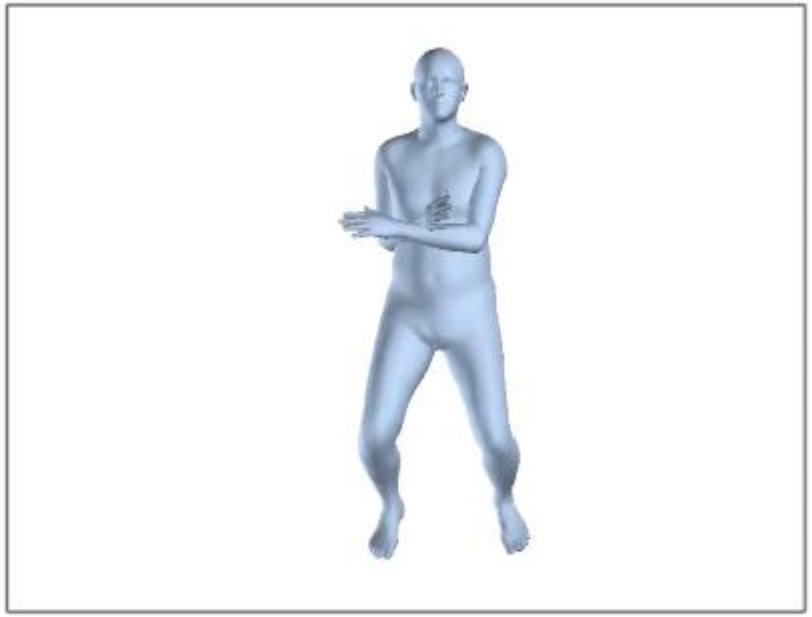
Video



Policy

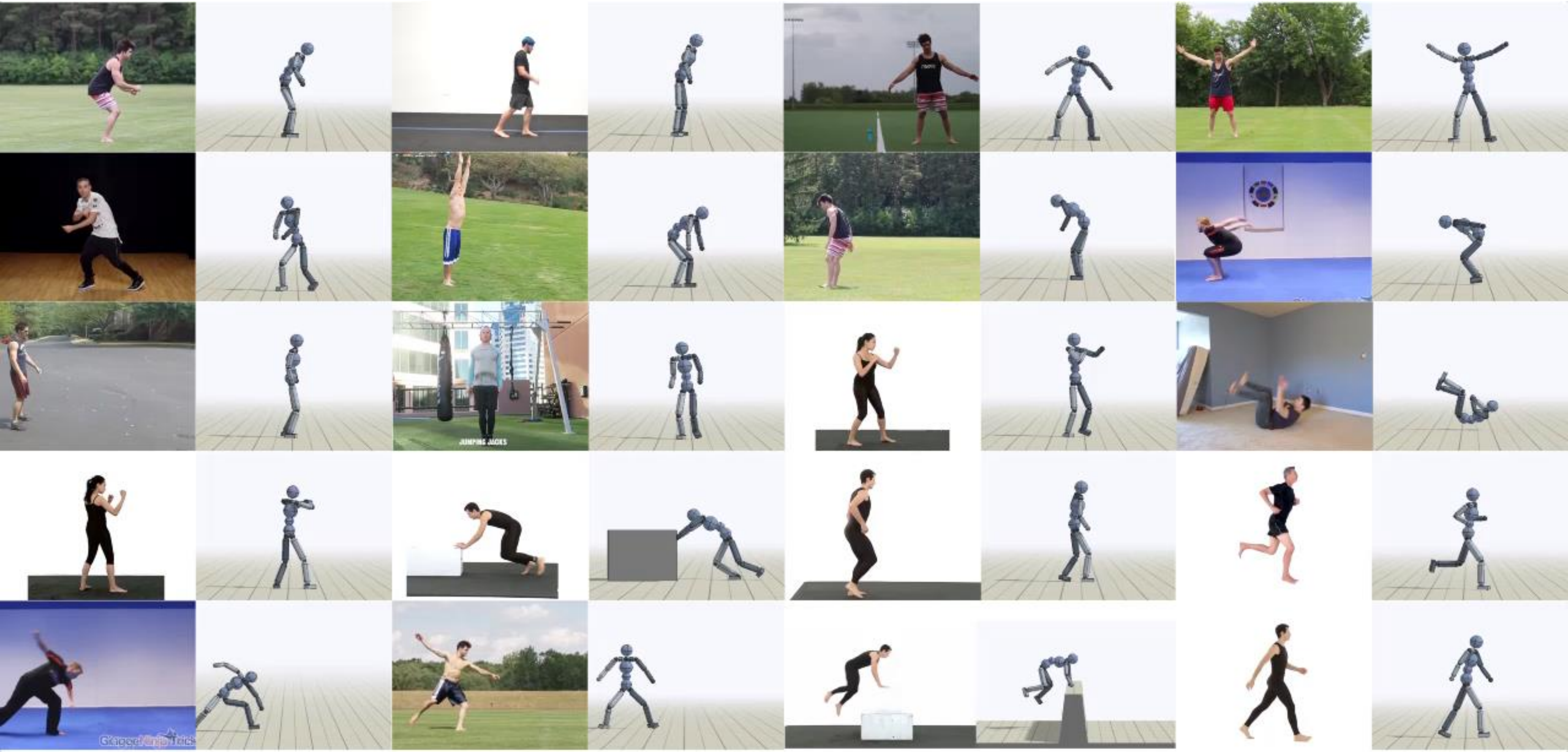
Failure Cases

Video



Recovered 3D Body

Policy


JUMPING JACKS