UNIVERSITÄT DES SAARLANDES

max planck institut informatik

Body Models I-1

Gerard Pons-Moll and Paul Swoboda

Max Planck Institute for Informatics

December 12, 2018

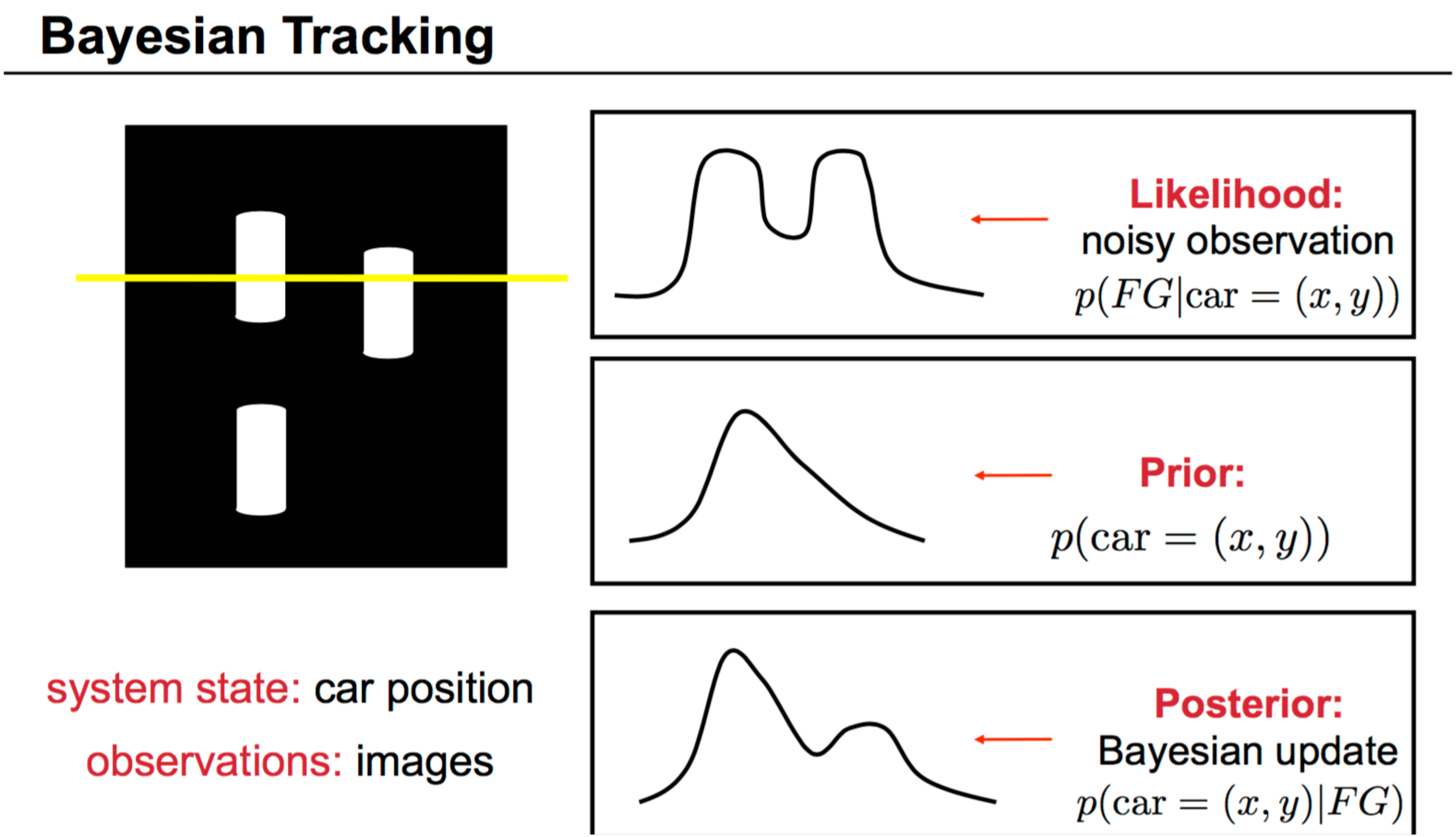
12.12.2018	**Body Models 1**
19.12.2018	Body Models 2
09.01.2019	Body Models 3
16.01.2019	Sampling and tracking
23.01.2019	Graphical models in computer vision
06.02.2019	Wrap up

Visual Inference of Humans

Why do humans require special attention ?

- At the centre of many applications
- Humans are complex with variation in articulation, shape, clothing and appearance

Bayesian Inference in Computer Vision



How to Model the Likelihood?

$$p(\mathbf{x}|\mathbf{z}) = \underbrace{p(\mathbf{z}|\mathbf{x})}_{\text{Likelihood}}\underbrace{p(\mathbf{x})}_{\text{Prior}}$$

Posterior

Likelihood

Prior

$\mathbf{x}$ State (variables of interest)

$\mathbf{z}$ Observations from sensors such as images, audio, IMU

How does this relate to discrete energy minimization?

$$\max_{\mathbf{x}\in X_V} p(\mathbf{x}|\mathbf{z}) = p(\mathbf{z}|\mathbf{x})p(\mathbf{x})$$

$$\downarrow$$

$$\min_{\mathbf{x}\in X_V} \psi(\mathbf{z},\mathbf{x}) + \phi(\mathbf{x})$$

$$\downarrow$$

$$\min_{\mathbf{x}\in X_V} \sum_{v\in V} \psi_v(\mathbf{z},\mathbf{x}_v) + \sum_{uv\in\mathcal{E}} \phi_{uv}(\mathbf{x}_u,\mathbf{x}_v)$$

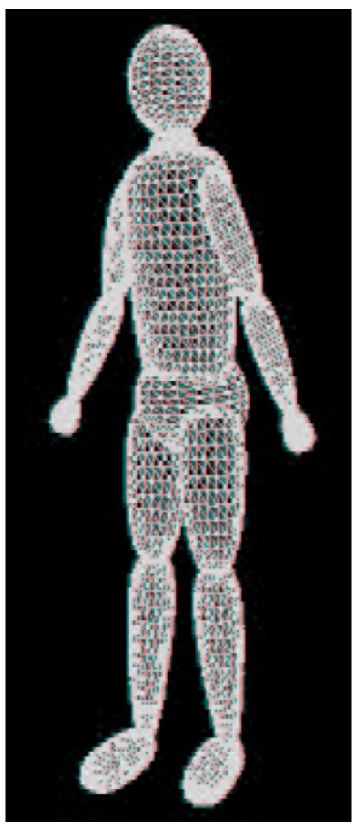
How to Model the Energy?

$$\min_{\mathbf{x}\in X_V} \sum_{v\in V} \psi_v(\mathbf{z}, \mathbf{x}_v) + \sum_{uv\in\mathcal{E}} \phi_{uv}(\mathbf{x}_u, \mathbf{x}_v)$$

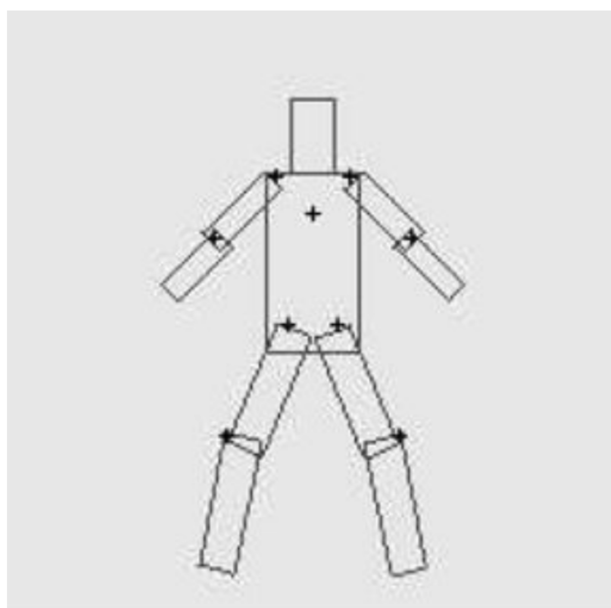
$\mathbf{z}$

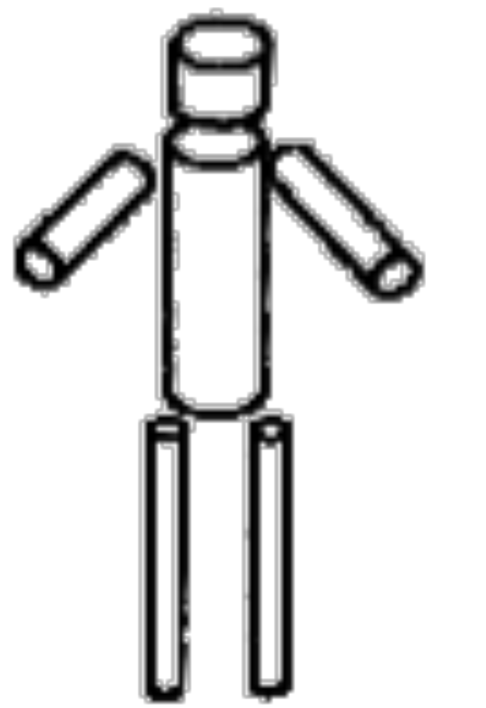
$\mathbf{x}$

What is the Right Level of Abstraction?



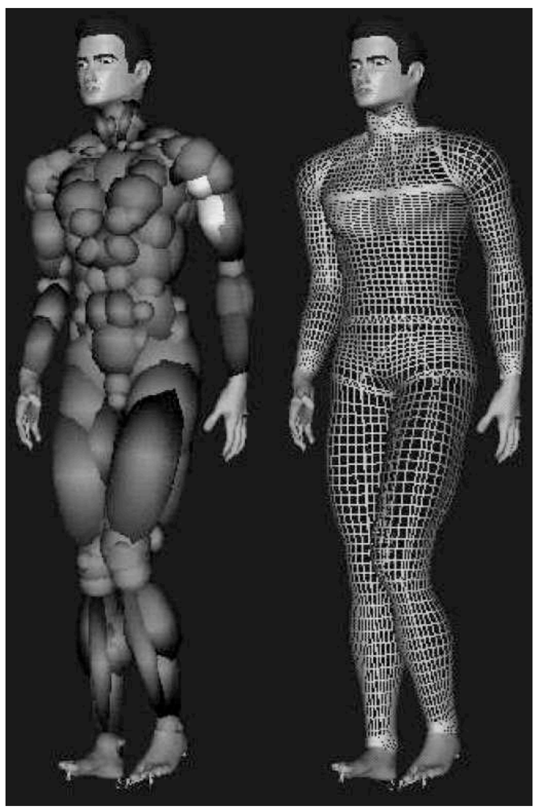
[Sminchisescu and Triggs ’03]





Marr and Nishihara ’78

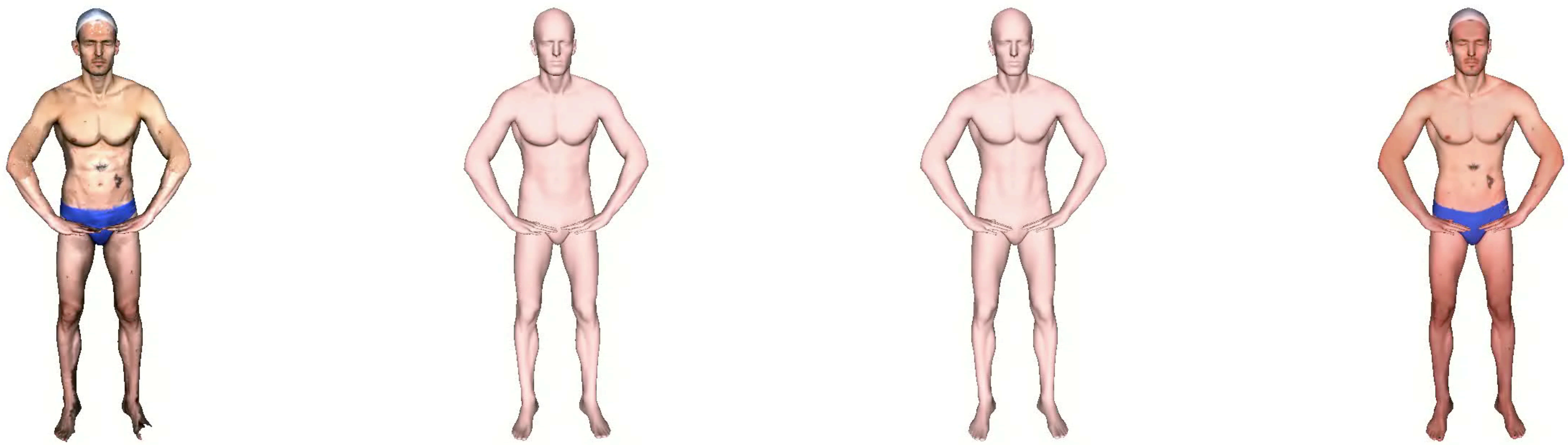
Nevatia and Binford ‘73



[Plänkers and Fua ’01]

Goal: Generative Model of People

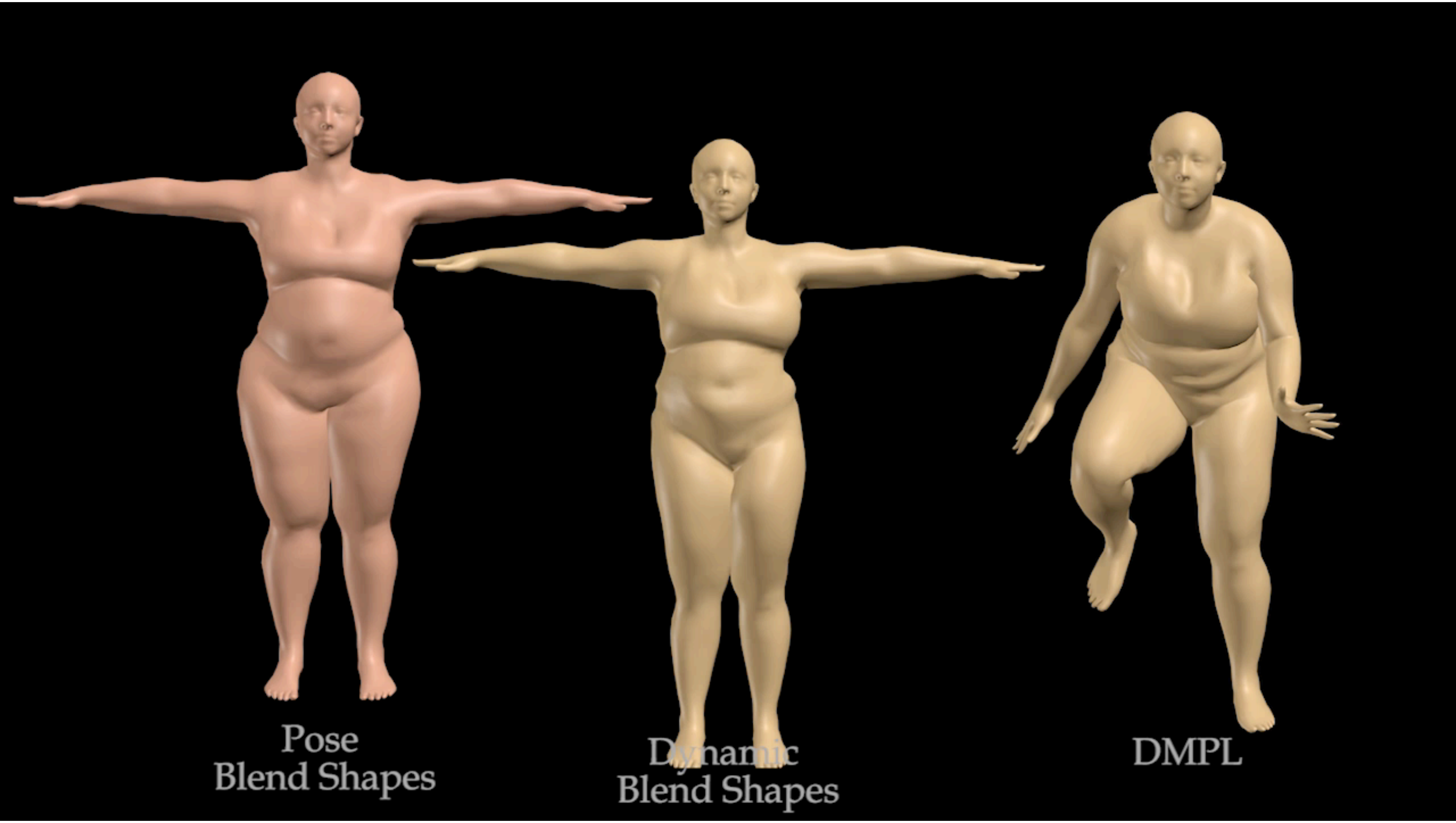
Body Model Function



$$M(\vec{\theta}, \vec{\beta}, \vec{u}; \Phi) \longrightarrow \text{3D mesh}$$

pose shape texture Hyper-parameters to learn

Modelling Soft-Tissue

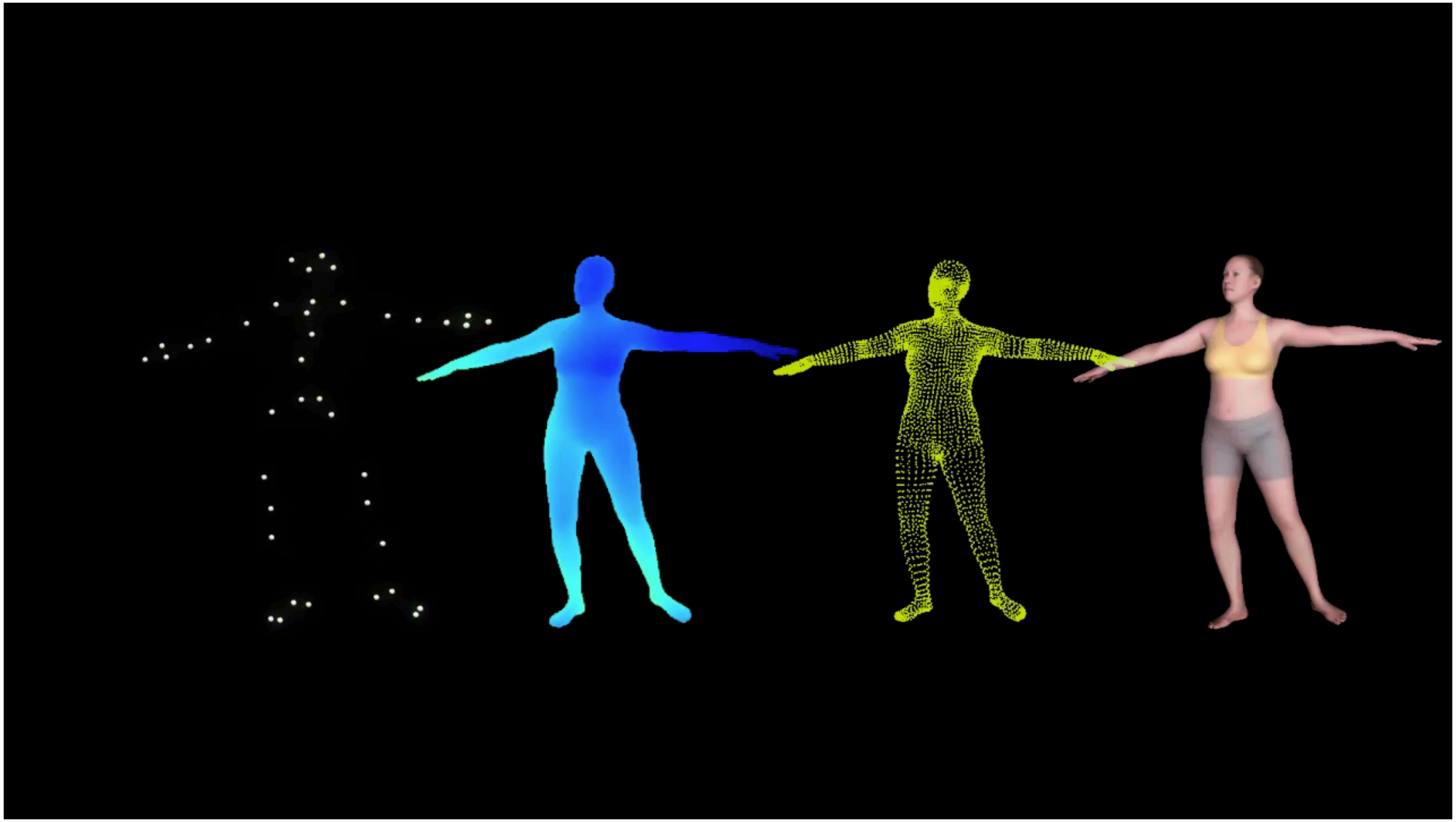


Pons-Moll et al. Siggraph 2015

Models of Clothing

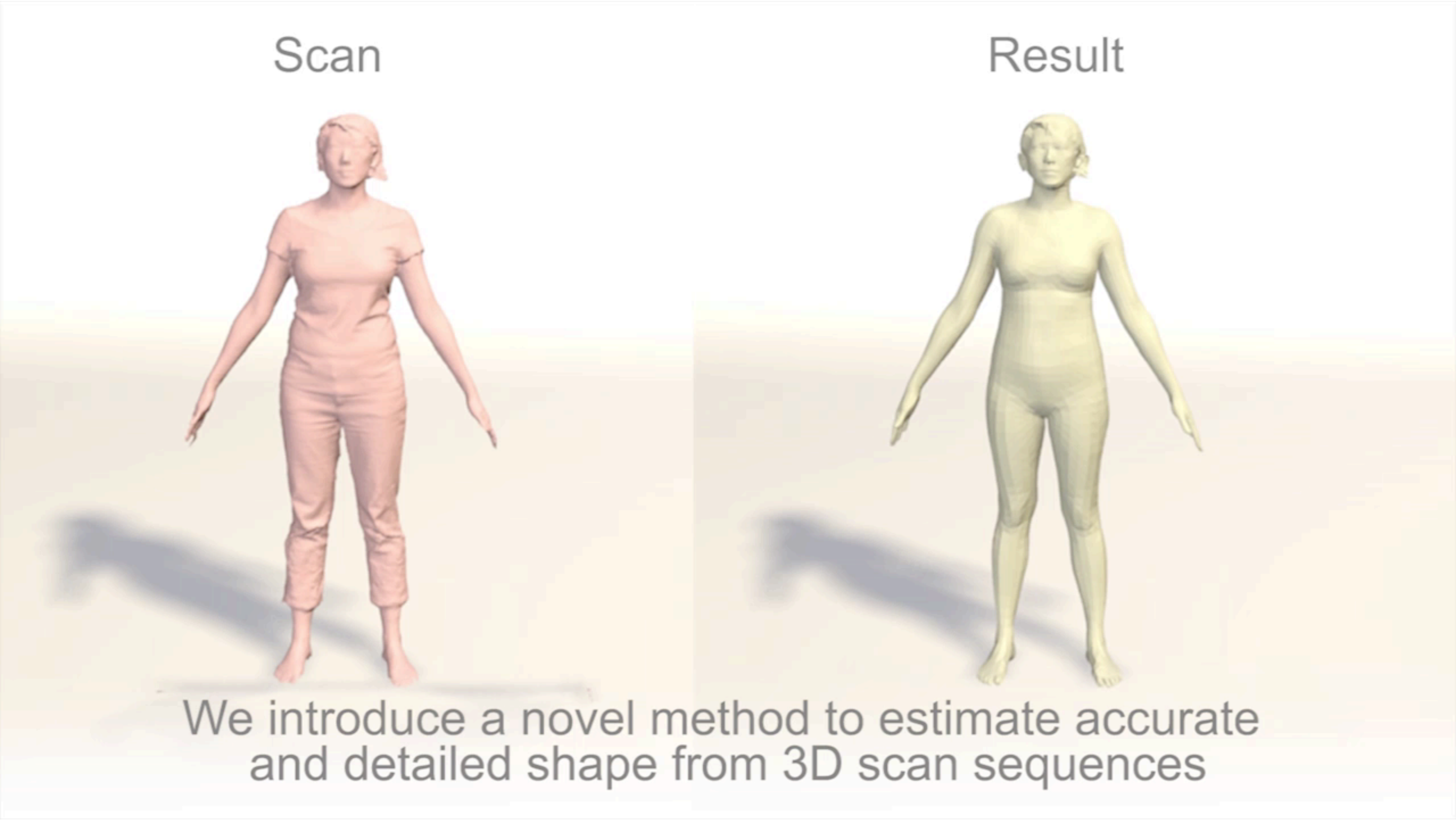
Pons-Moll et al. Siggraph 2017

Analysis by Synthesis



We can render, mocap markers, depth maps, point clouds, meshes, images, video, IMUs, measurements, … *and fit this to data or use it for training.*

Estimating Shape Under Clothing



Zhang et al. CVPR 2017

Sparse Inertial Poser

Automatic 3D Human Pose Estimation from Sparse IMUs

Supplementary material

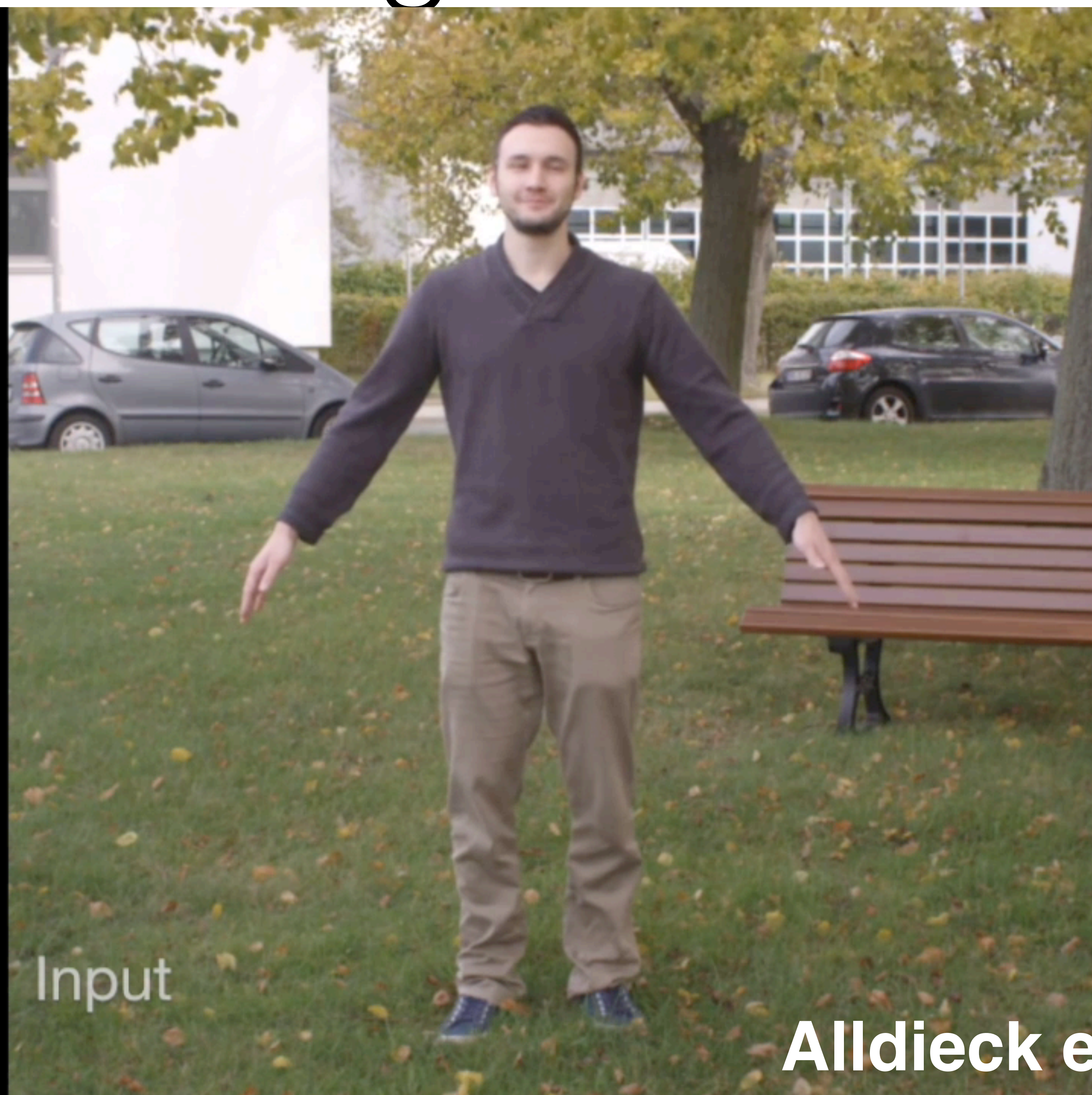
Eurographics'17
Paper ID 1112

von Marcard et al. Eurographics 2017 (best paper award)

Fitting the Model to RGB-D

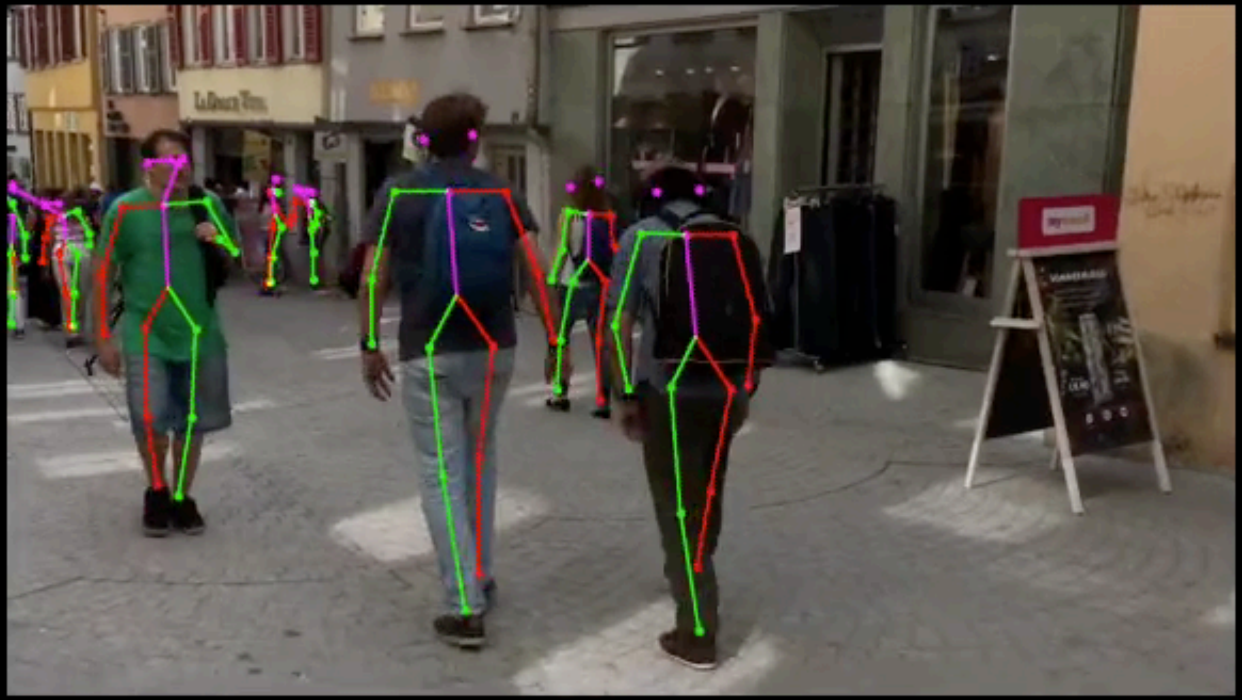
Tao et al. CVPR'18

3D reconstruction from a Single Video

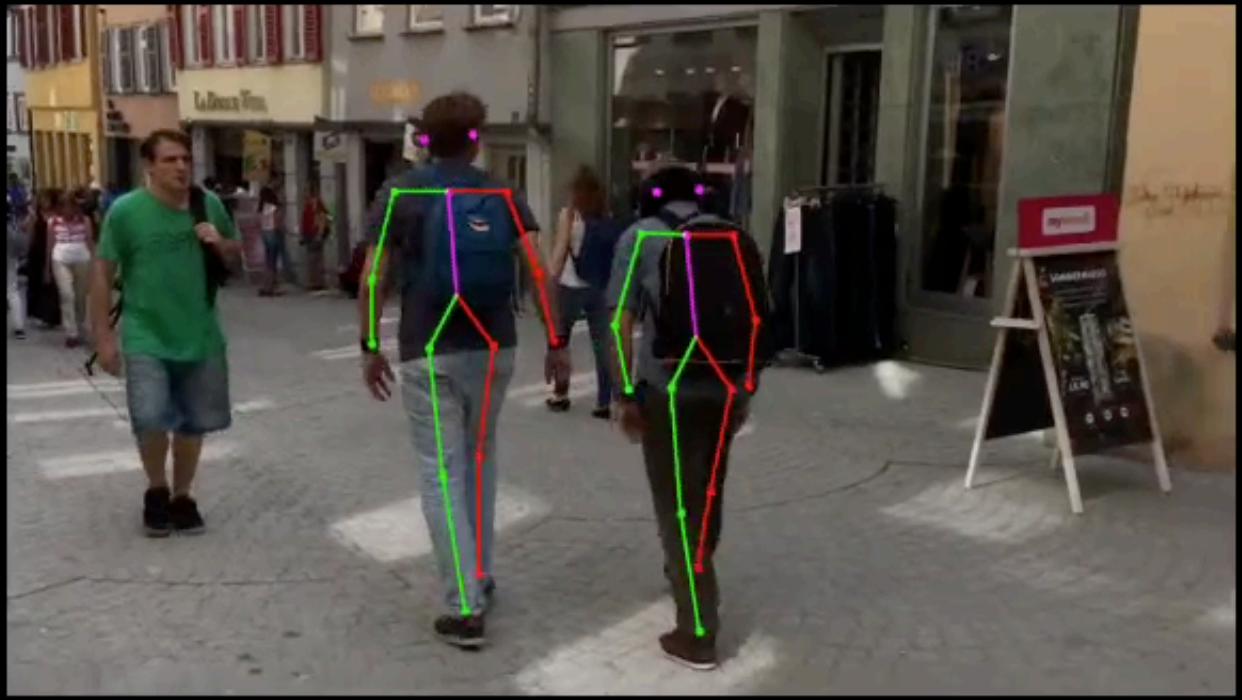


Alldieck et al. CVPR'18

Identify People



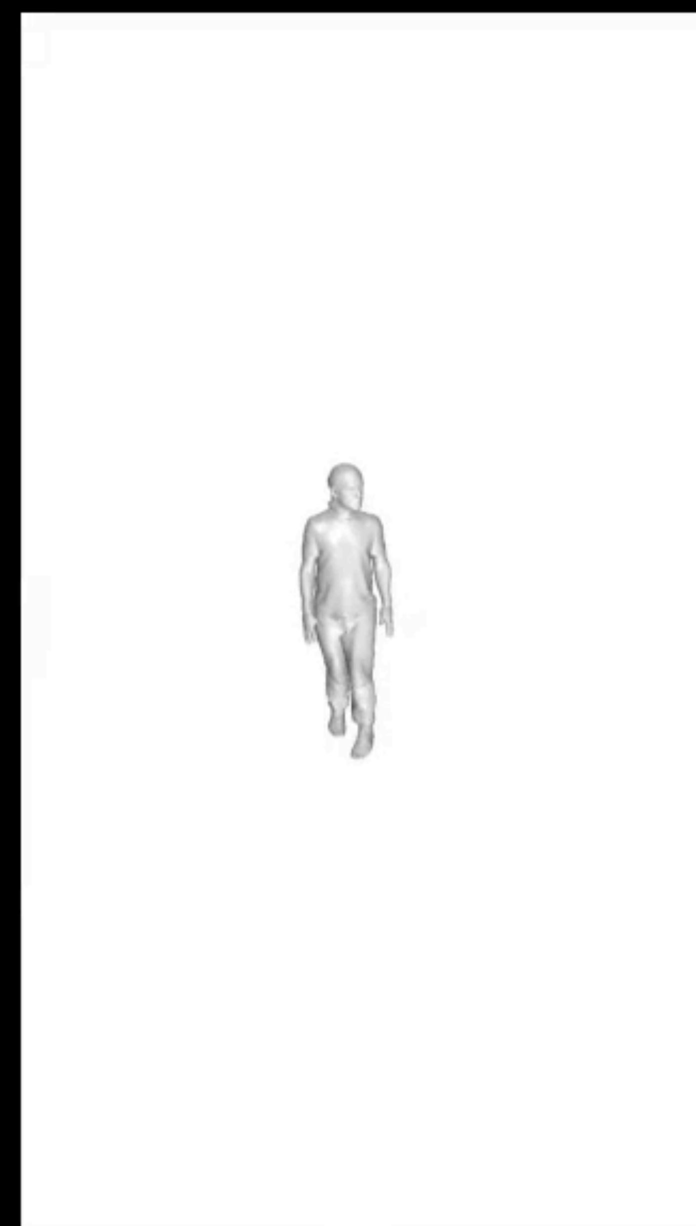
All 2D Poses



Assigned 2D Poses

3D reconstructions from Video and IMU

Projected 3D Pose



Animated 3D Pose

3D Human Pose and Shape from a Single Image

Omran et al. **3DV Best student paper**

Face, Hands, Body

Monocular Total Capture Results

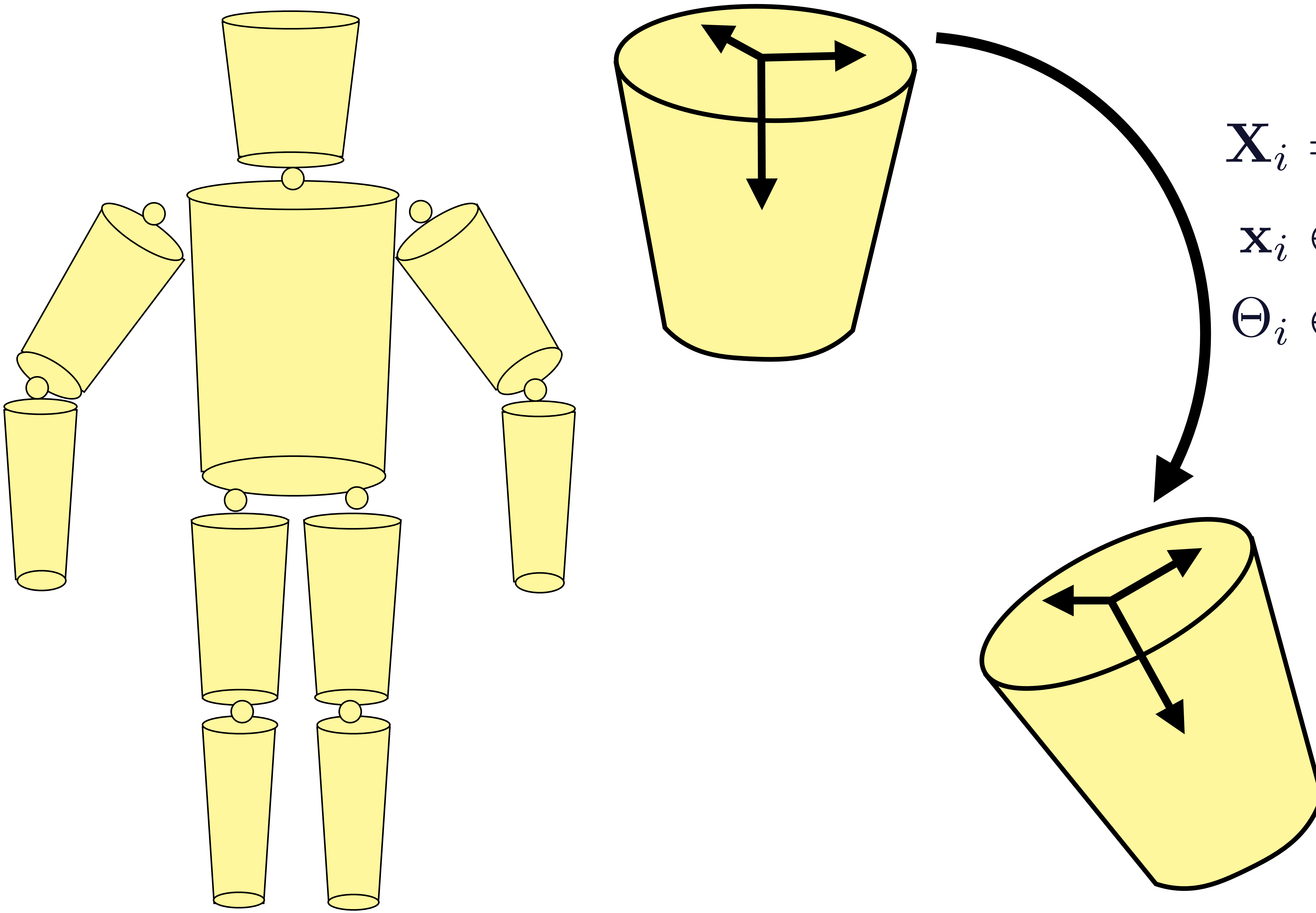
Input: a monocular video
Output: total body (face, hands, body, feet) motion

Building the generative model:

- How do we represent the body ?
- How do we learn the model from data ?

If we learned the generative model:

- How do we extract information about people from different sensors ?
- What inference techniques should we use ?

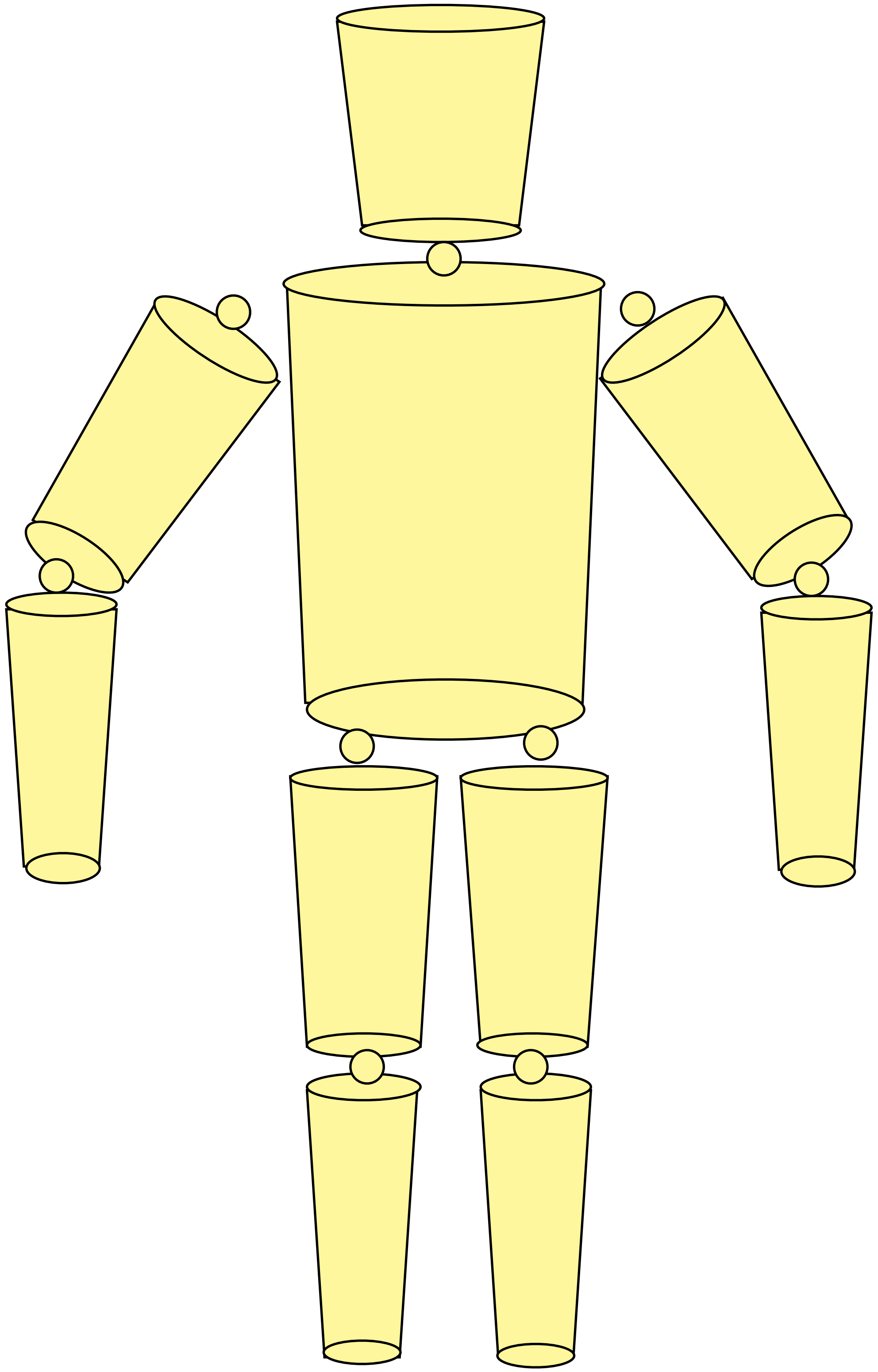


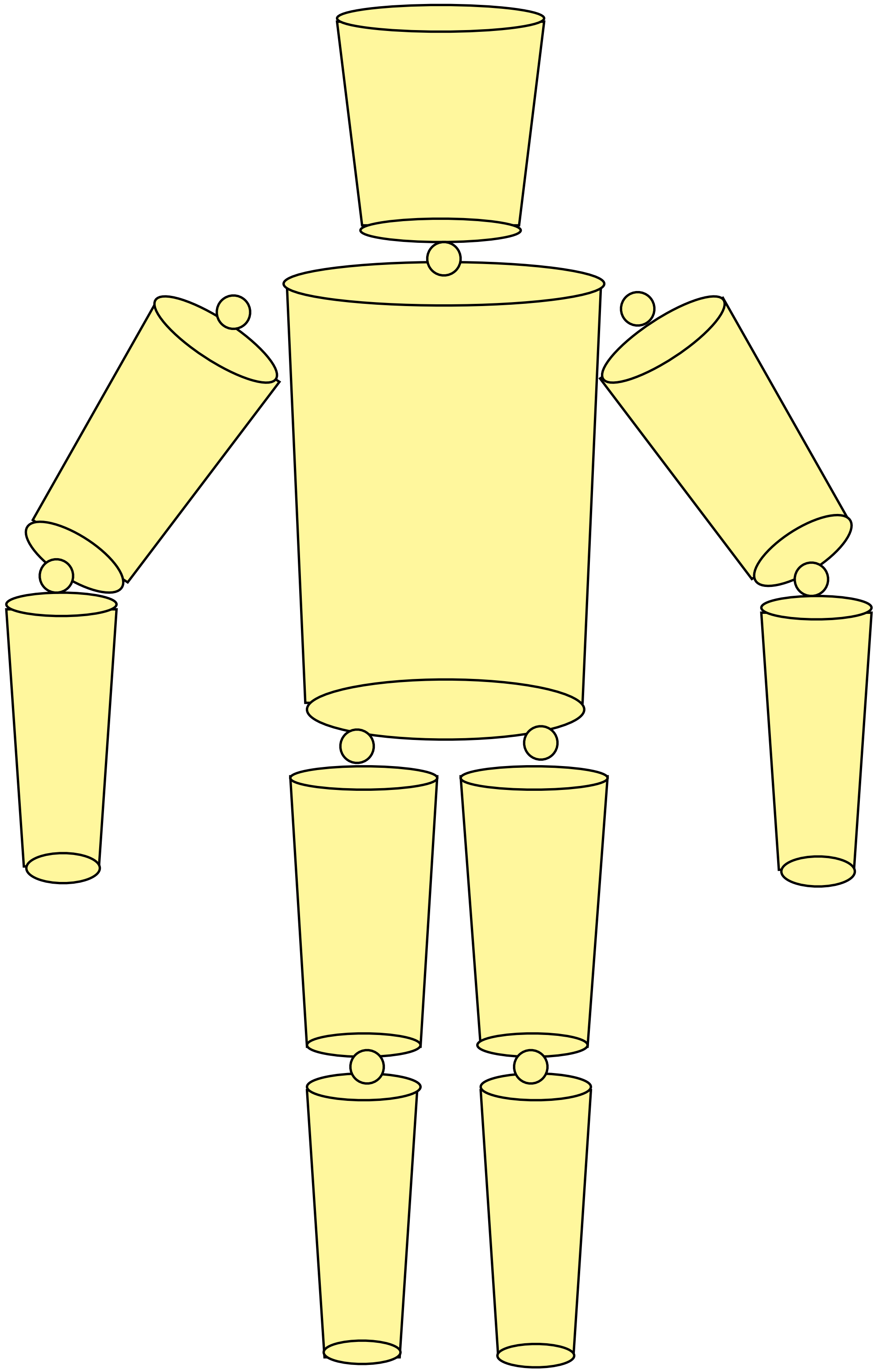
$$\mathbf{X}_i = (\mathbf{x}_i, \Theta_i)$$

$$\mathbf{x}_i \in \mathbb{R}^3$$

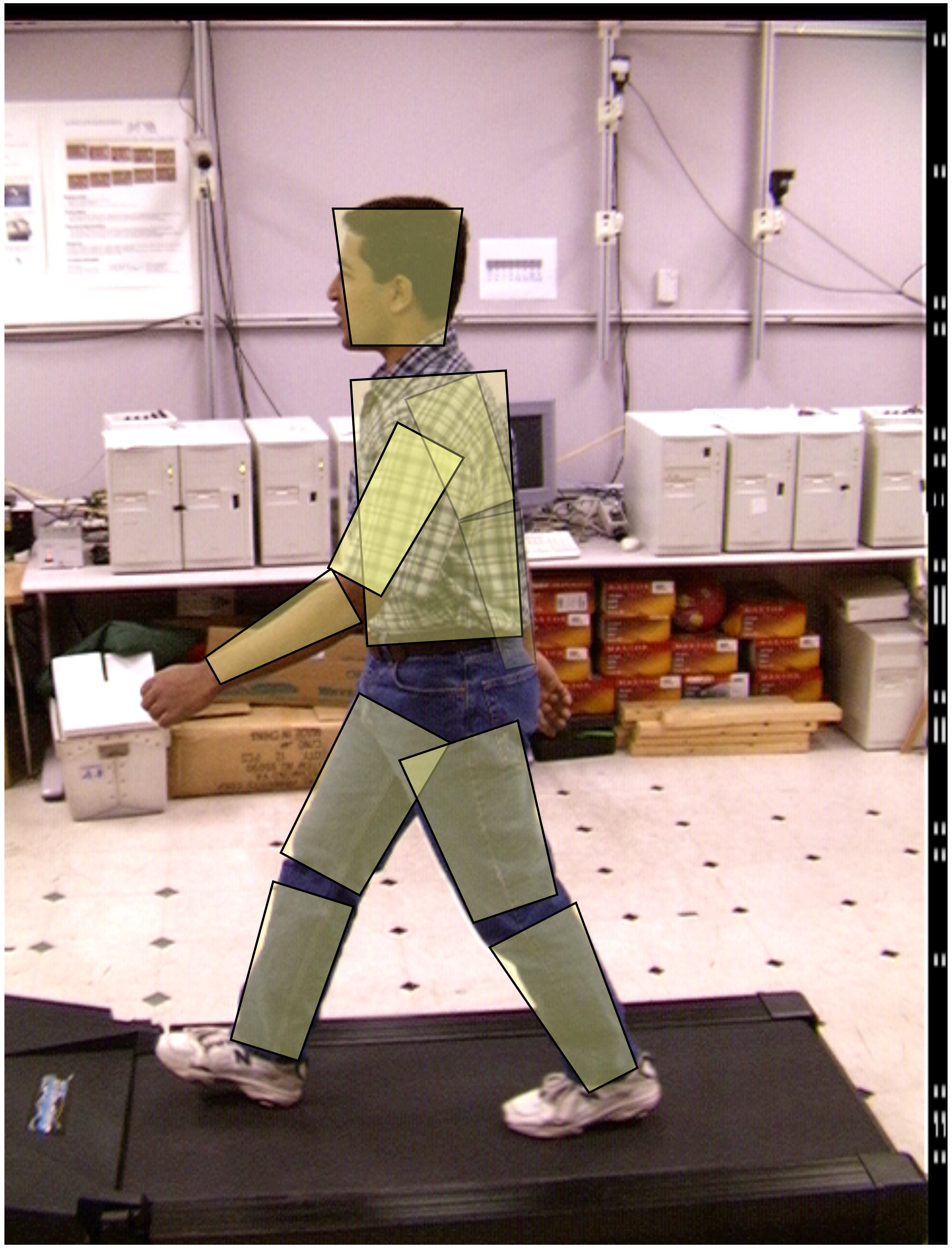
$$\Theta_i \in \mathbf{SO}(3)$$

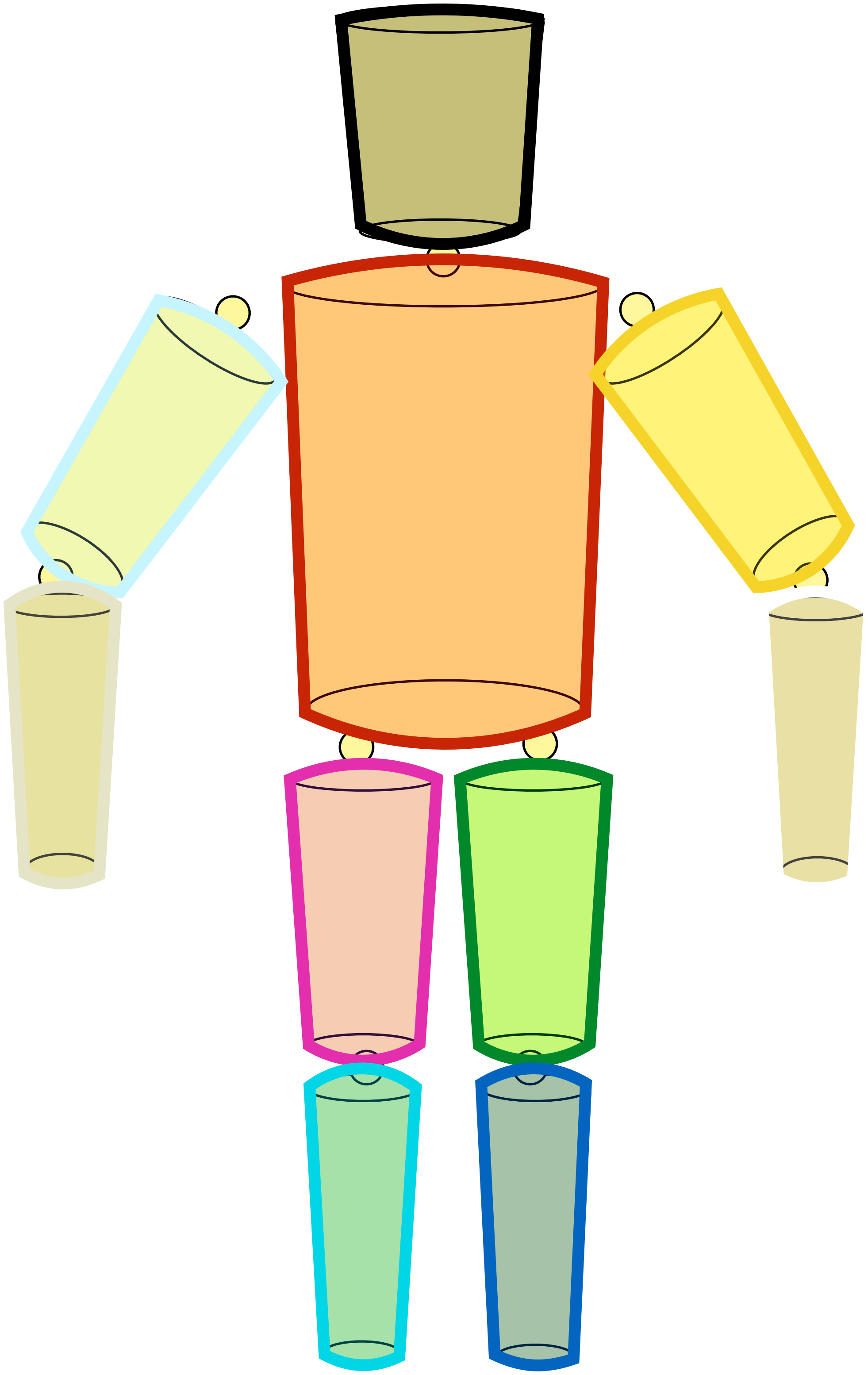
Attractive People: Assembling Loose-Limbed Models using Non-parametric Belief Propagation, Sigal et al.



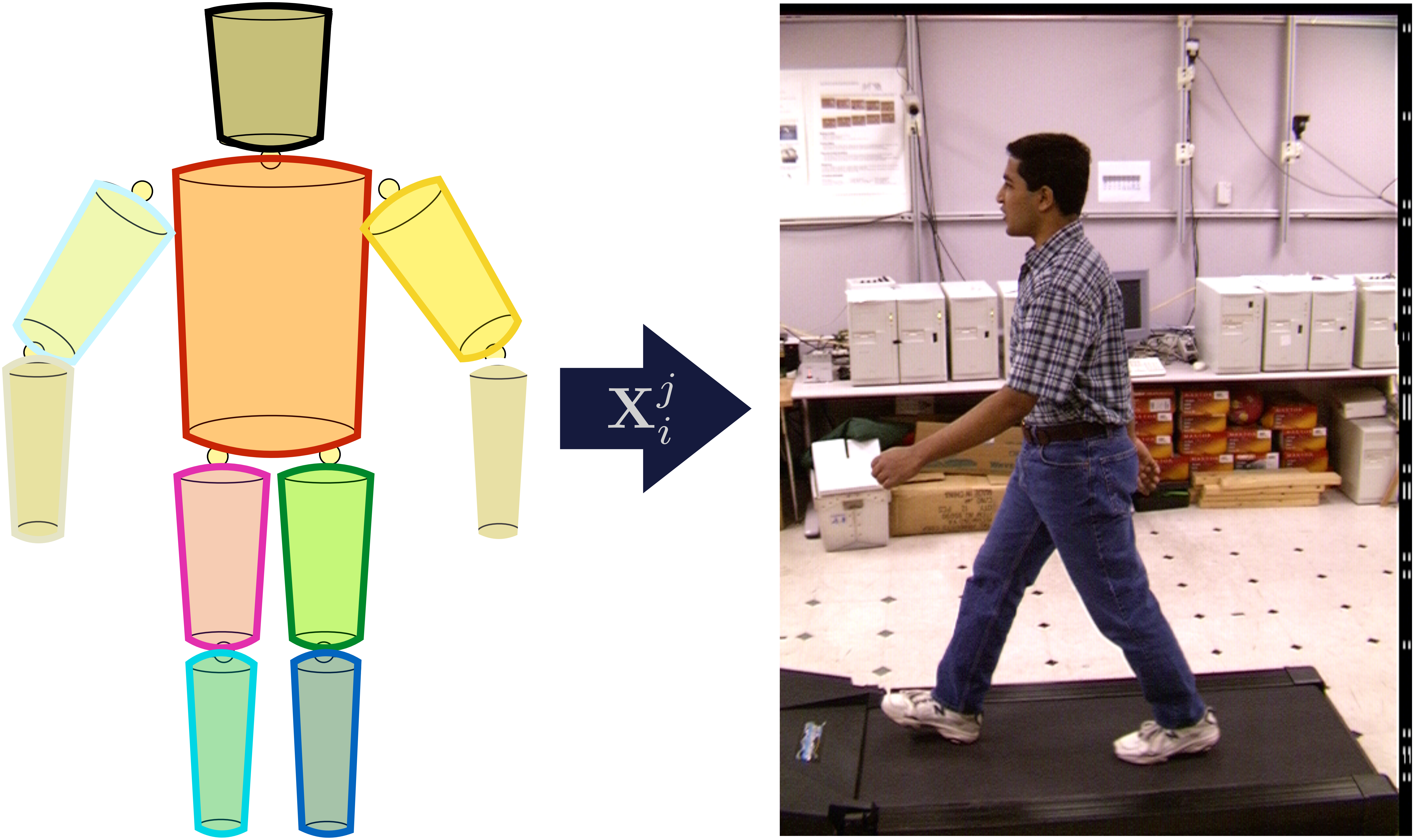


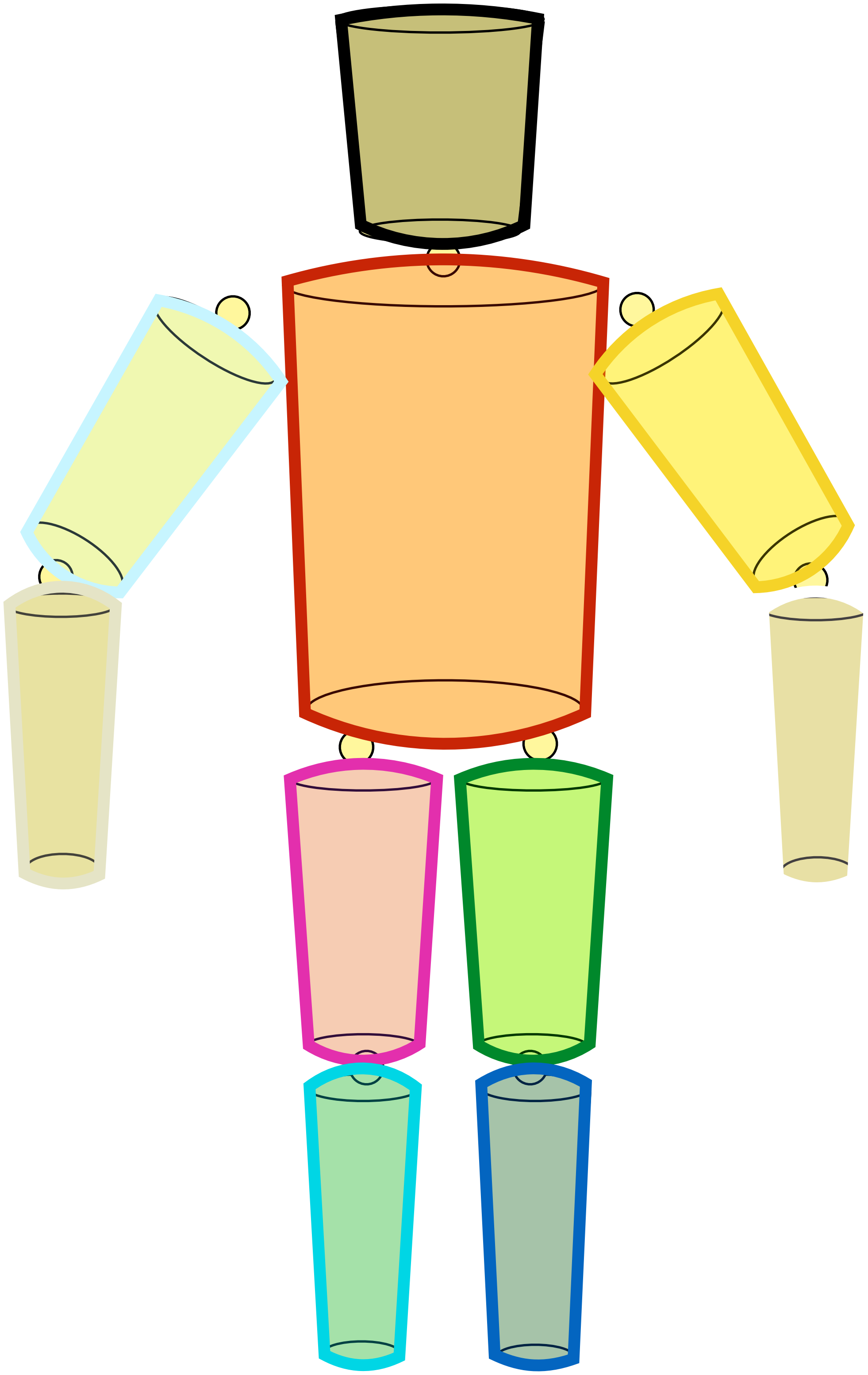
Goal



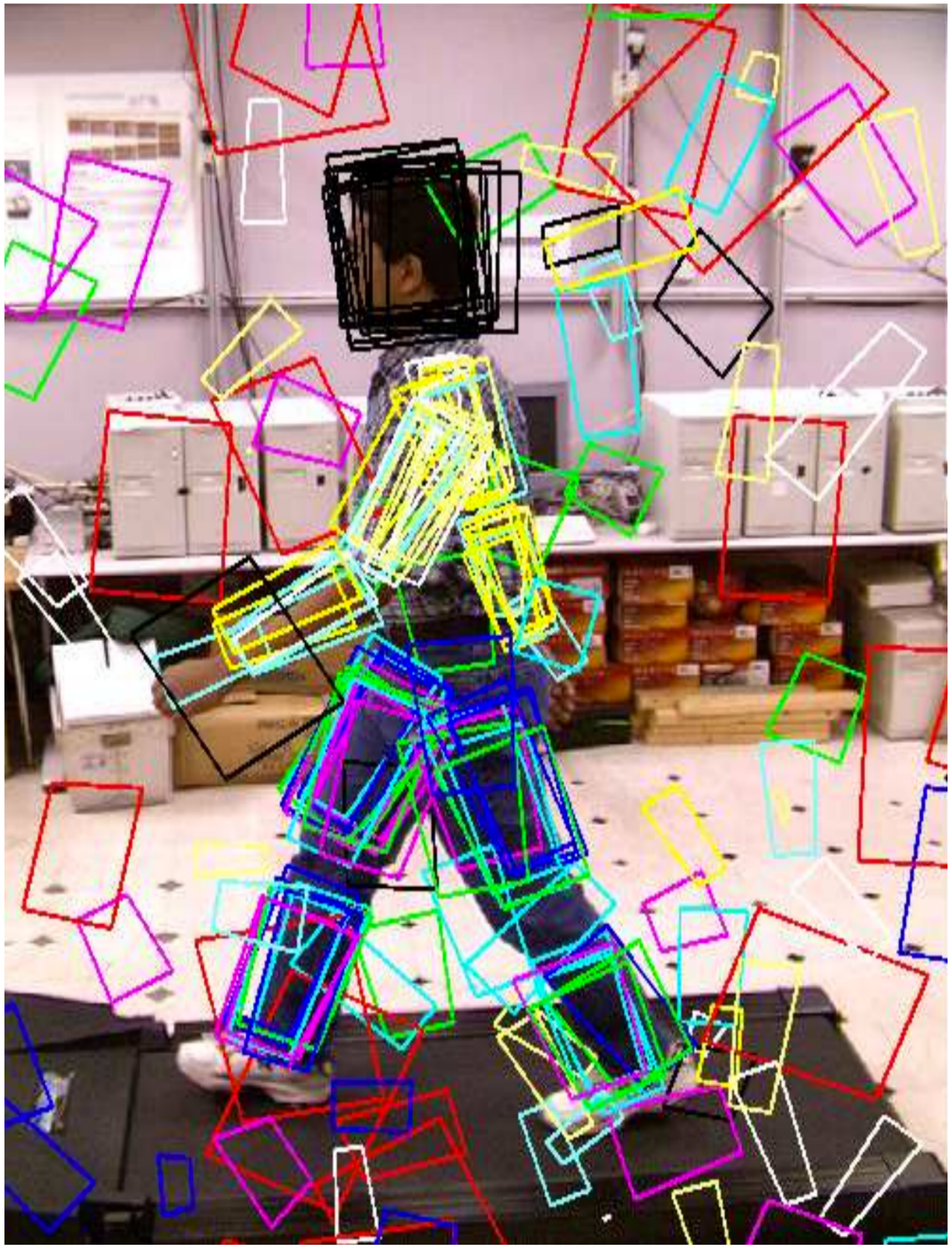


$$\mathbf{X}_i = \arg\min_{\mathbf{X}_i} \phi_i(I, \mathbf{X}_i)$$


$\mathbf{x}_i^j$



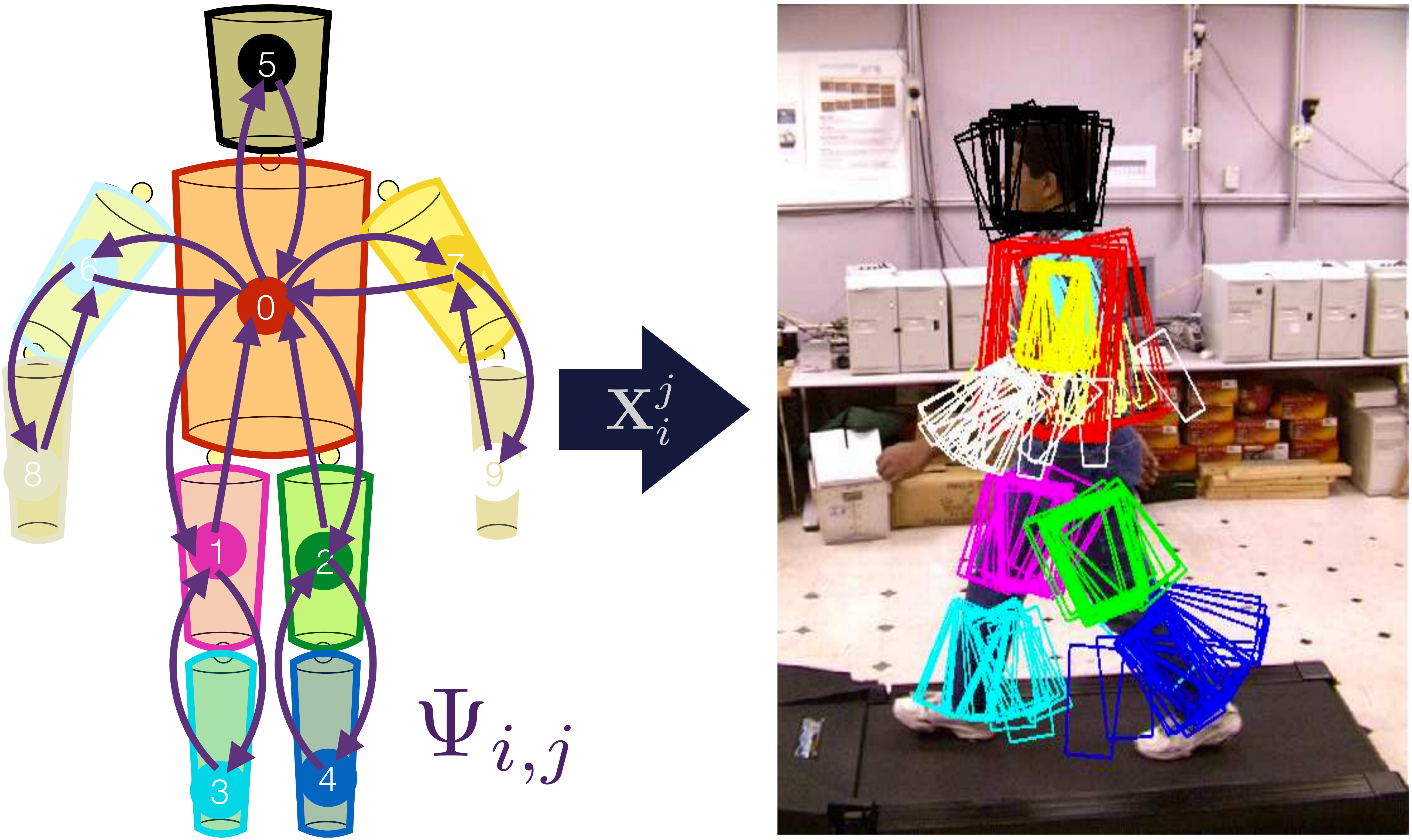
$\mathbf{x}_i^j$



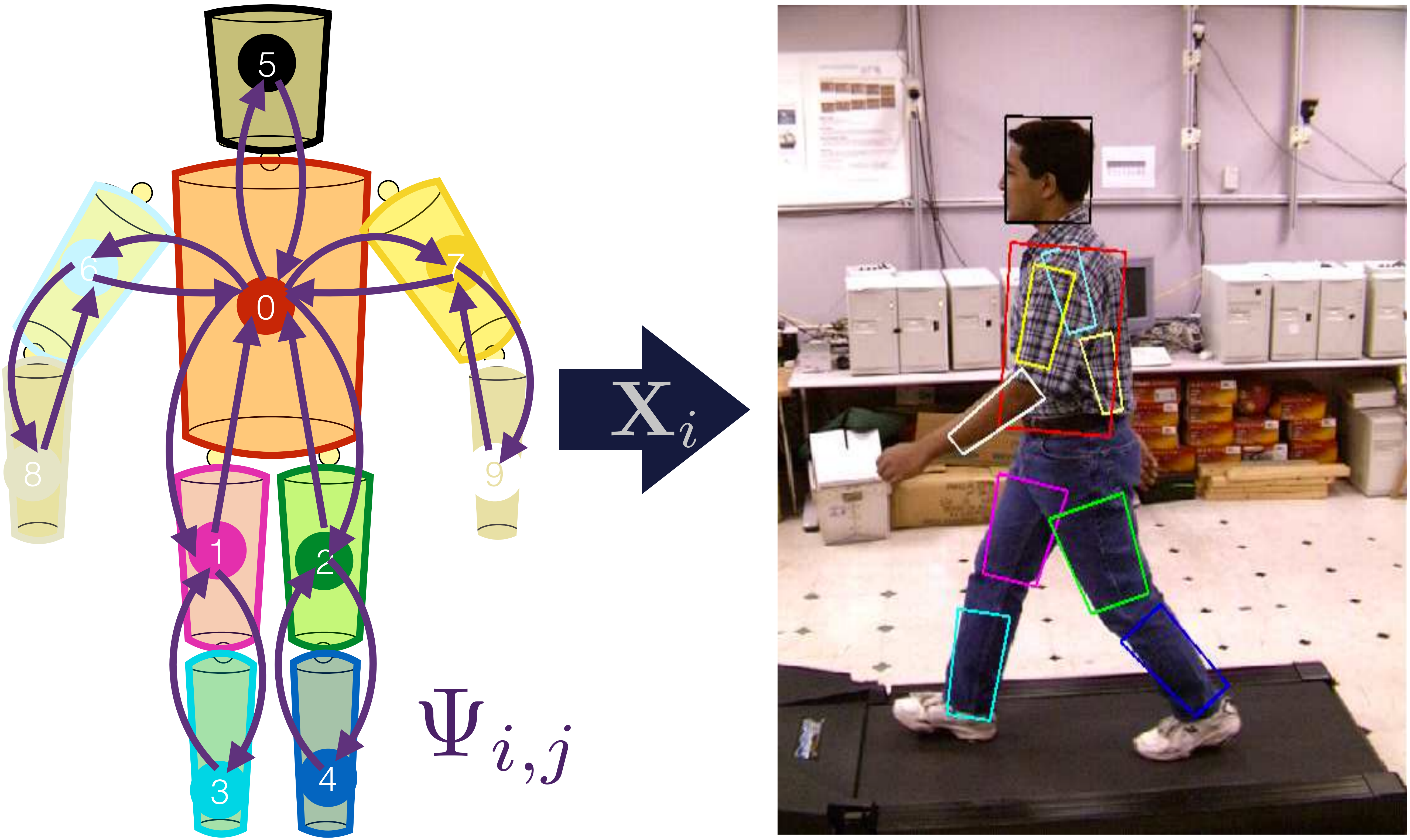
Can we do something better?

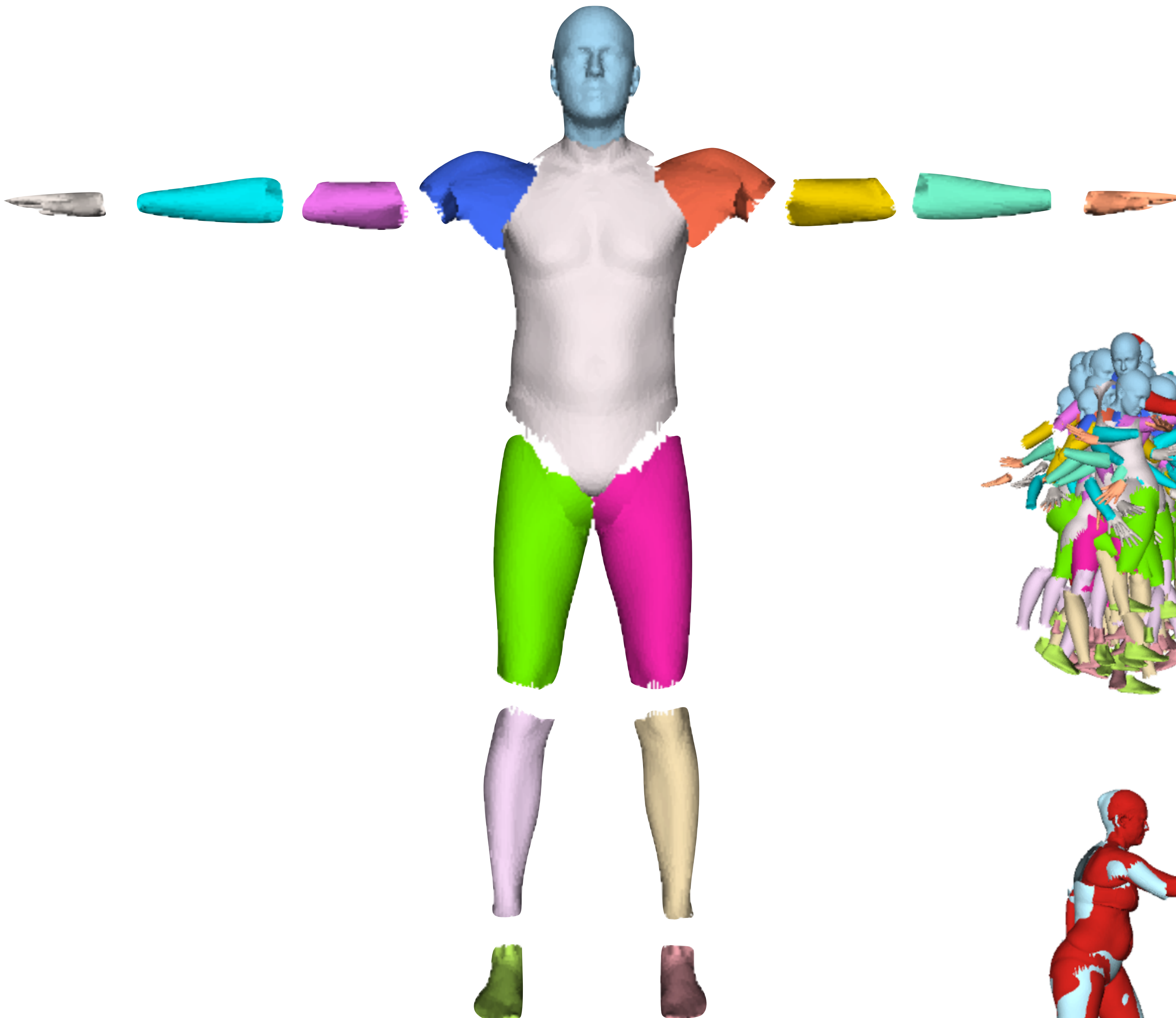
Can we do something better?

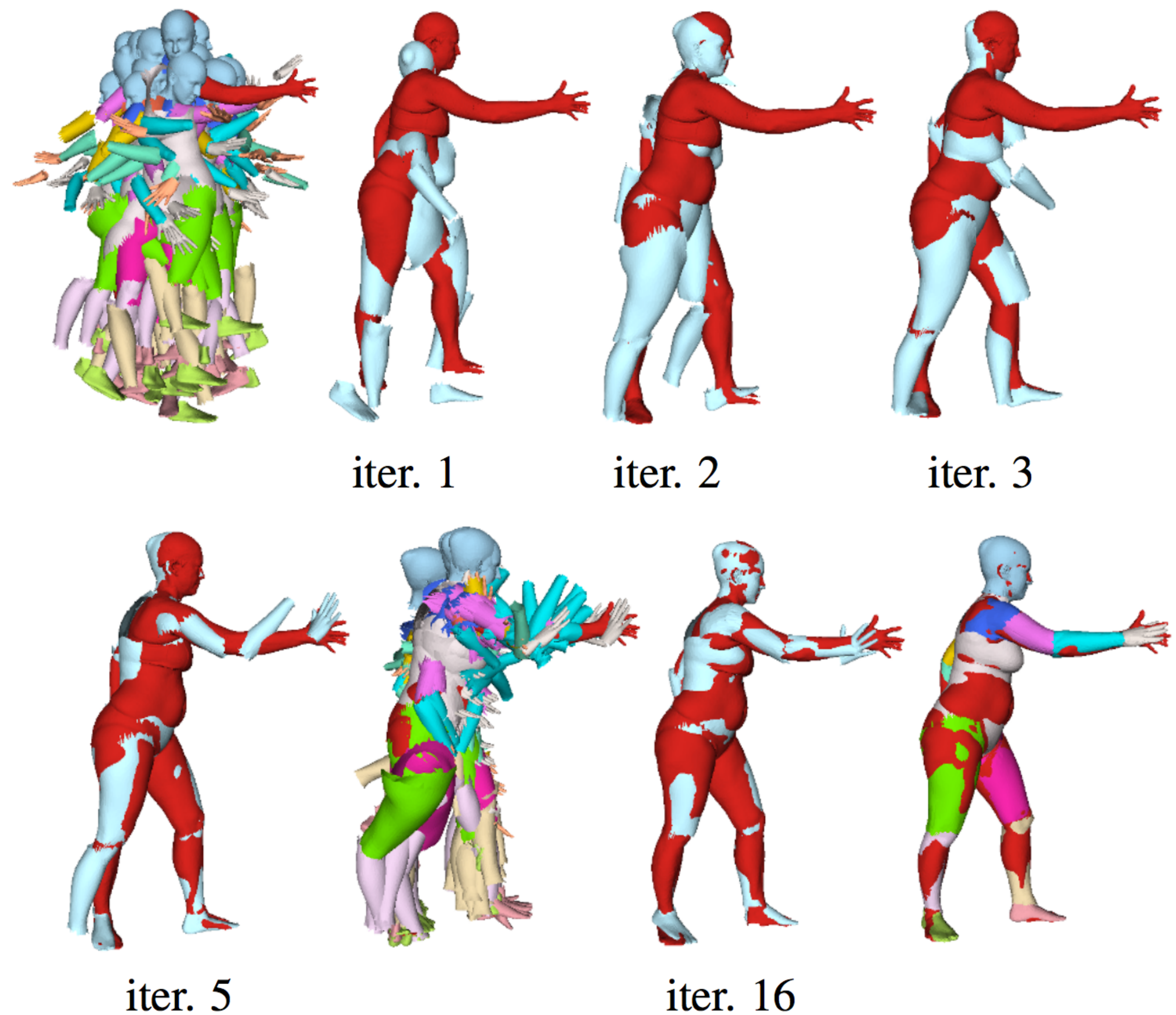
… with something that you already know?



5
6
7
0
8
9
1
2
3
4
$\Psi_{i,j}$
$\mathbf{x}_i^j$


5
6
0
7
8
9
1
2
3
4
$\Psi_{i,j}$
$\mathbf{X}_i$





The Stitched Puppet: A Graphical Model of 3D Human Shape and Pose, Zuffi and Black

Questions before we start

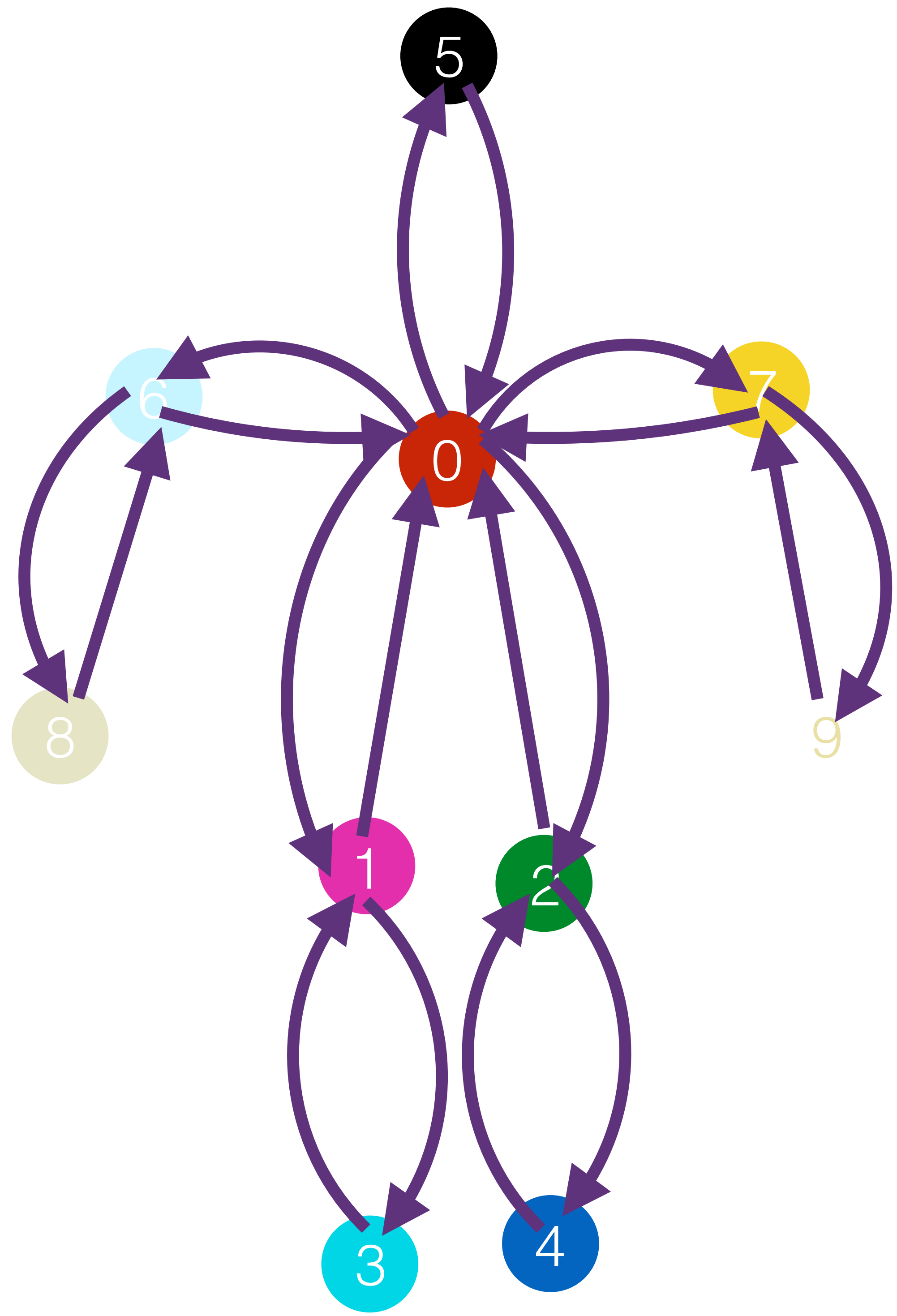
- What is a good unary ϕ_i for the pose estimation problem?
- What is a good pair-wise term $\Psi_{i,j}$?

Holistic Models

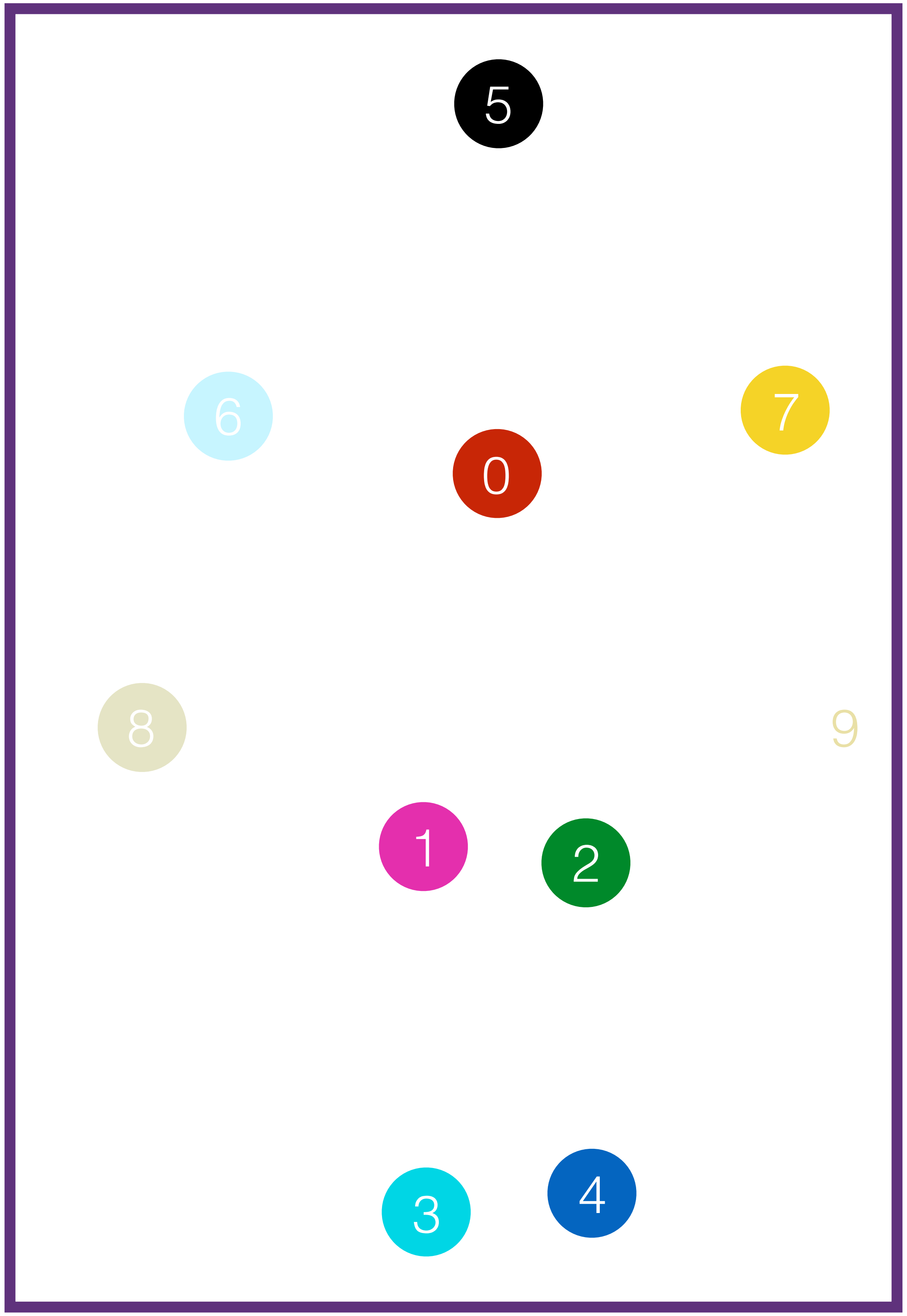
- With these ingredients we could attack the problem of pose estimation, but…

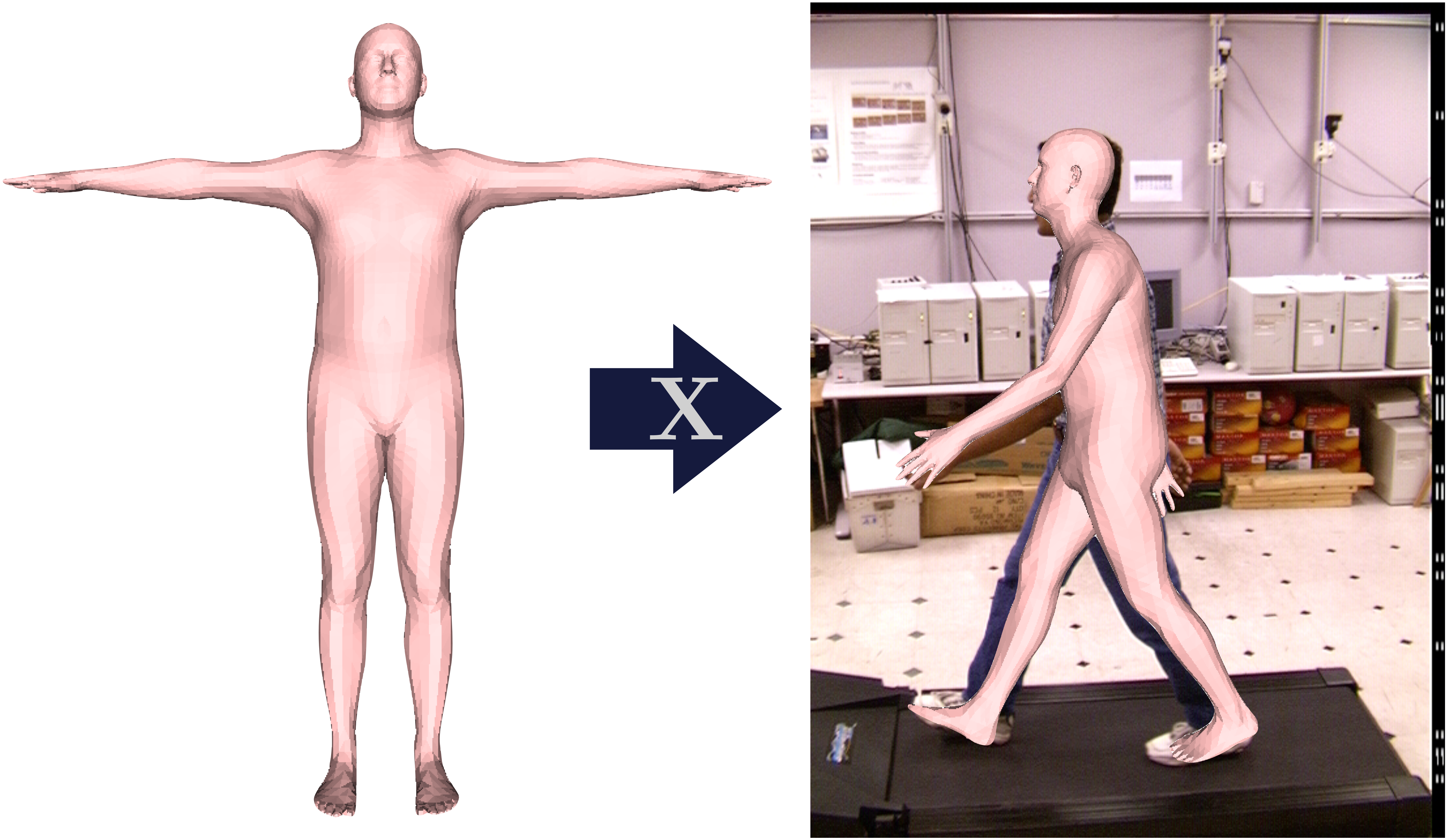
- Can we use another representation?

Holistic Models



Why not all at once?



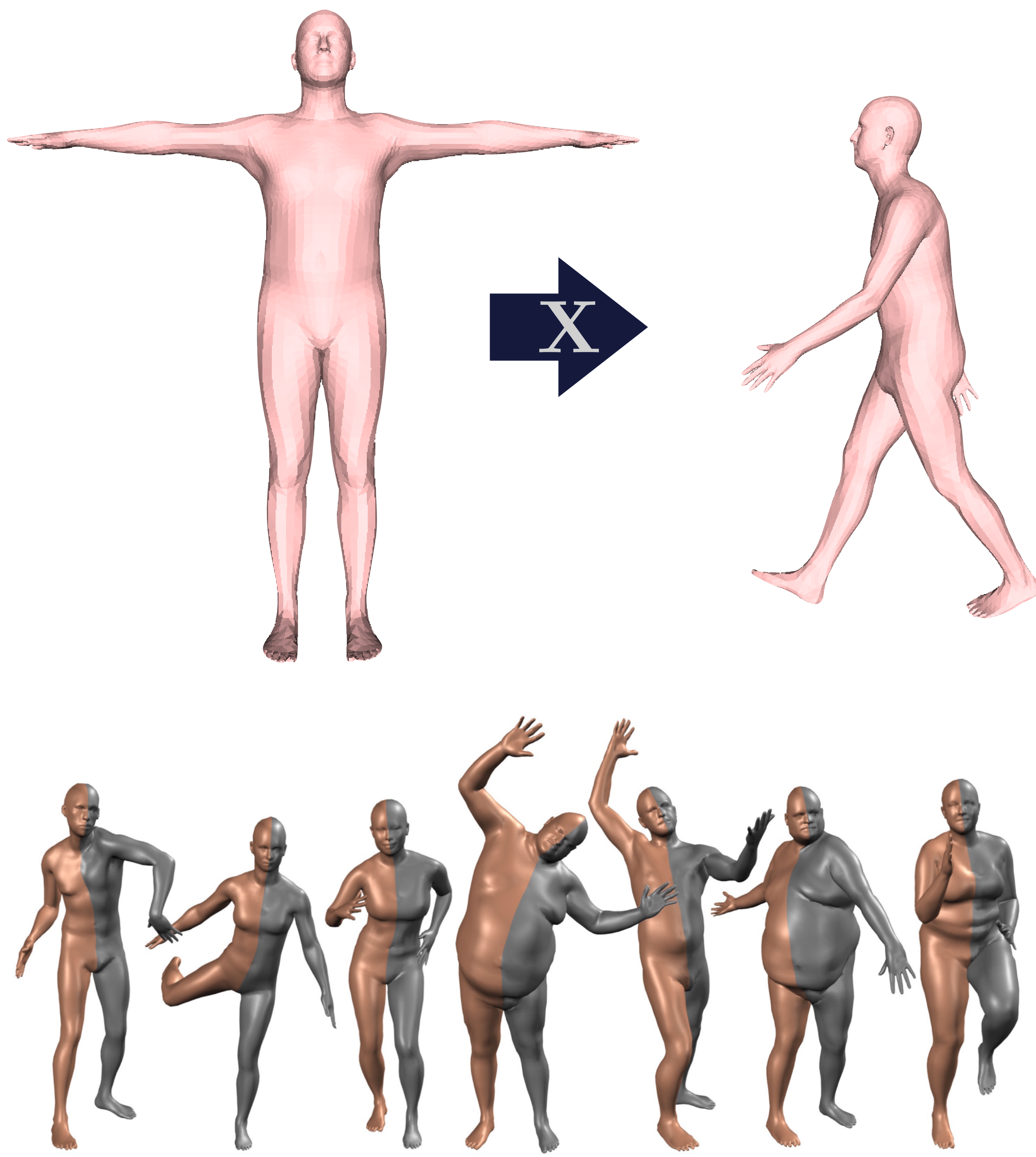


SMPL: A Skinned Multi-Person Linear Model, Loper et al.

Part-based models VS holistic models

- Part-based
  - Efficient
  - Fast exploration of parameter space
- Holistic
  - Part relation **imposed** by model, not optimised
  - Realistic solutions (up to the model's realism)

Part-based models VS holistic models



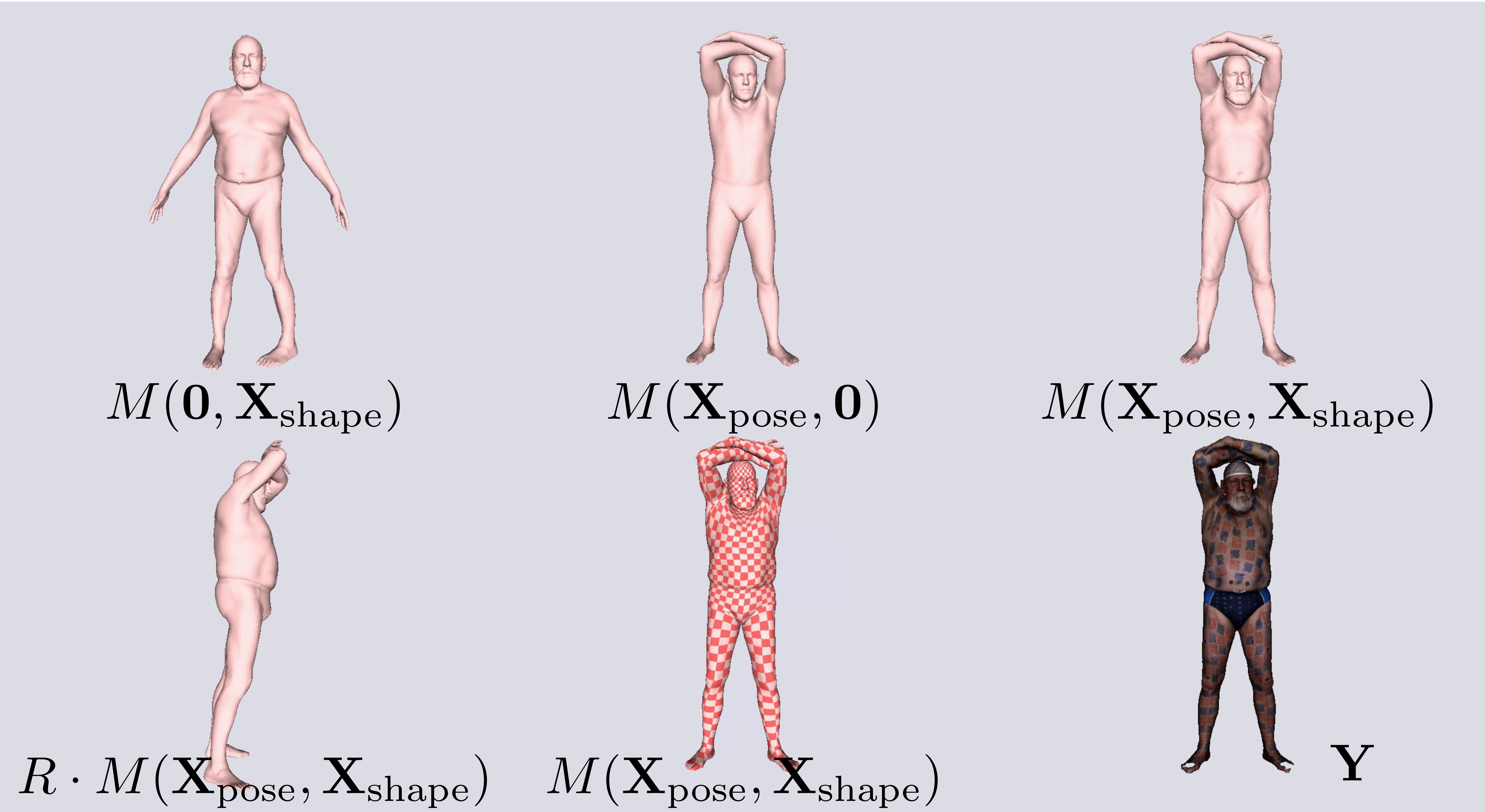
- Holistic
  - Part relation **imposed** by model, not optimised
  - Realistic solutions (up to the model's realism)

SMPL: A Skinned Multi-Person Linear Model, Loper et al.

3D Holistic Body Models

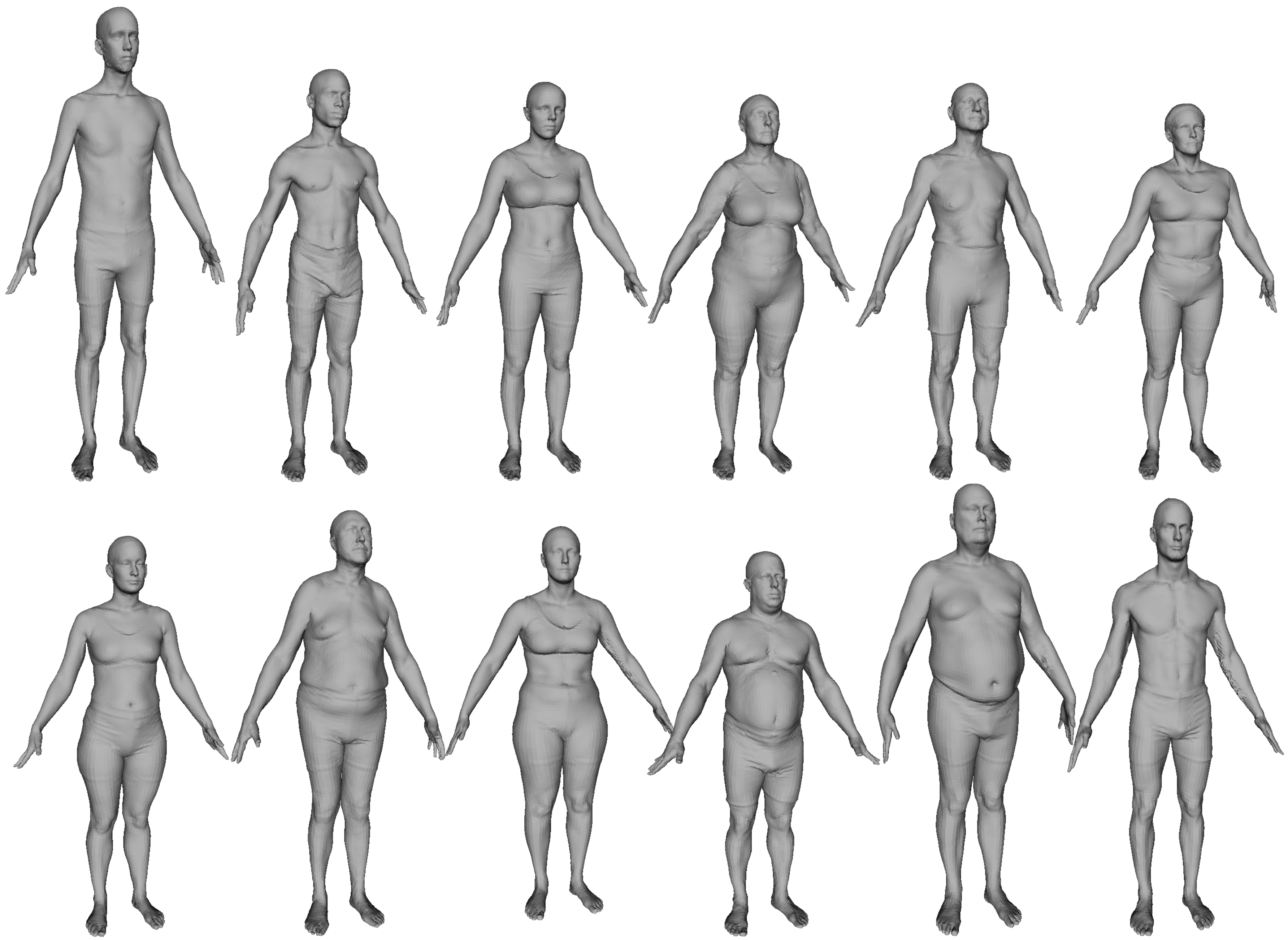
$$M(\mathbf{X}) \rightarrow \mathbb{R}^3$$

Transform parameters into geometry in 3D

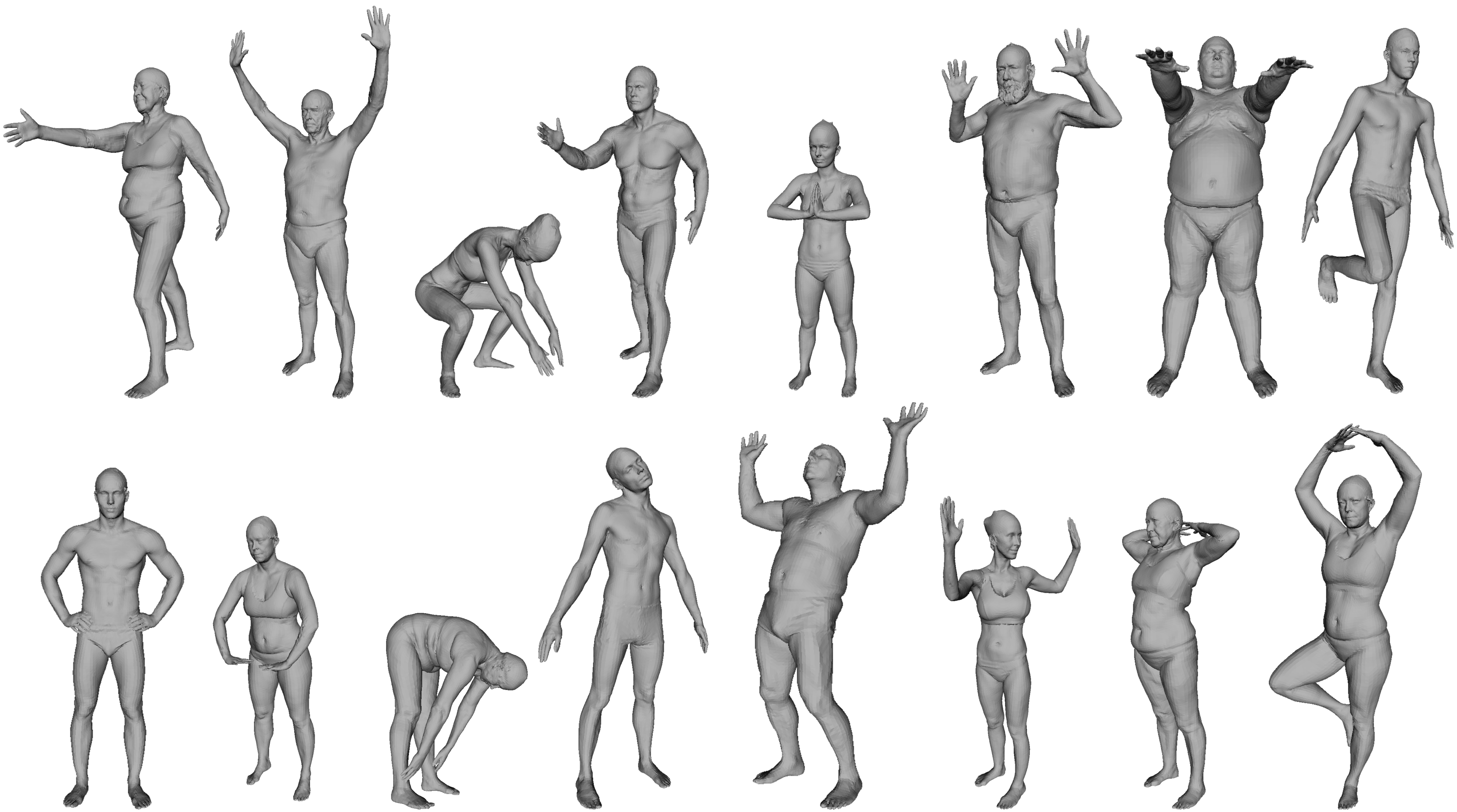


$$\mathbf{X} = \{\mathbf{X}_{\text{pose}}, \mathbf{X}_{\text{shape}}\}$$

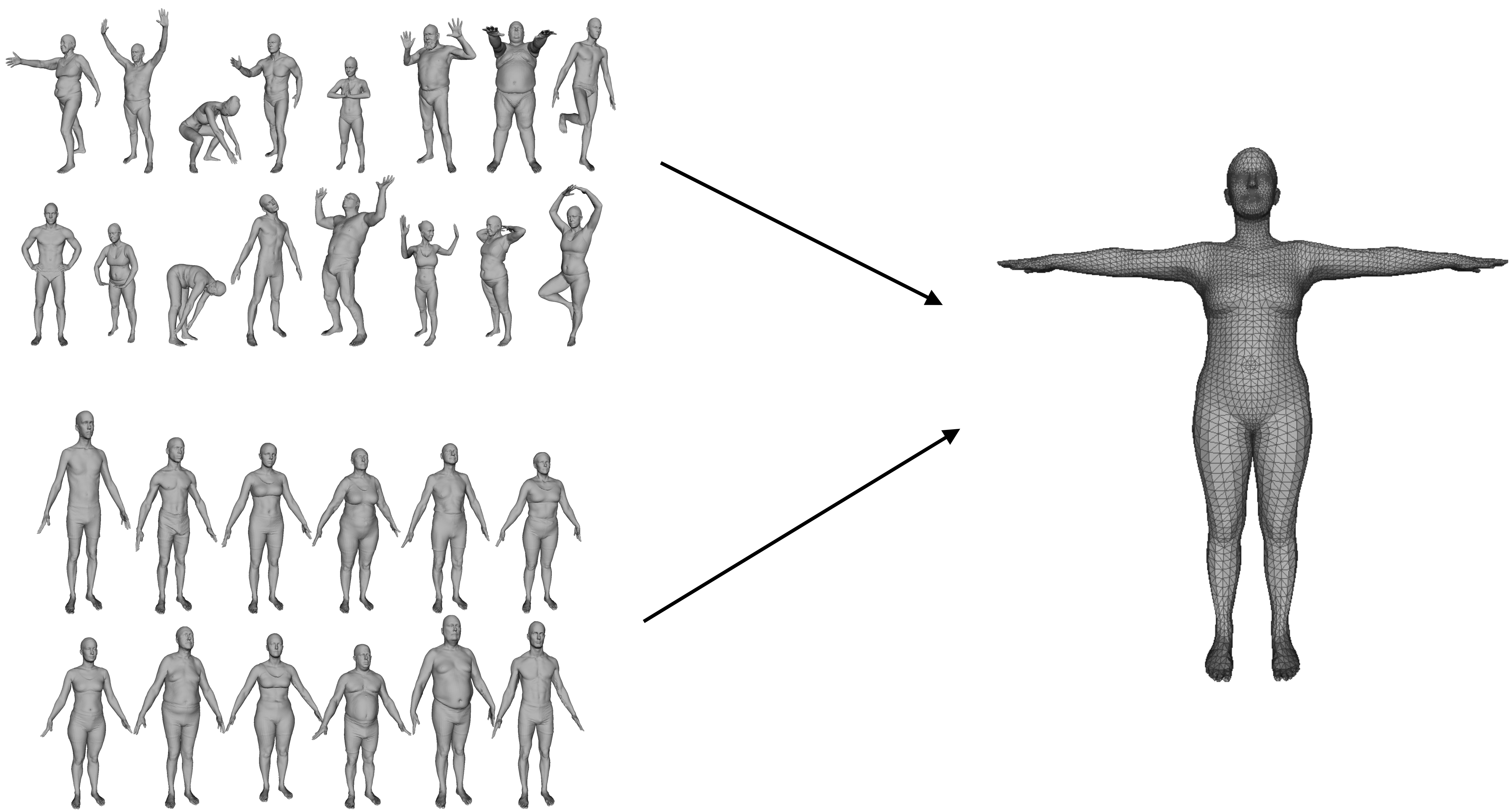
Scan a Lot of People

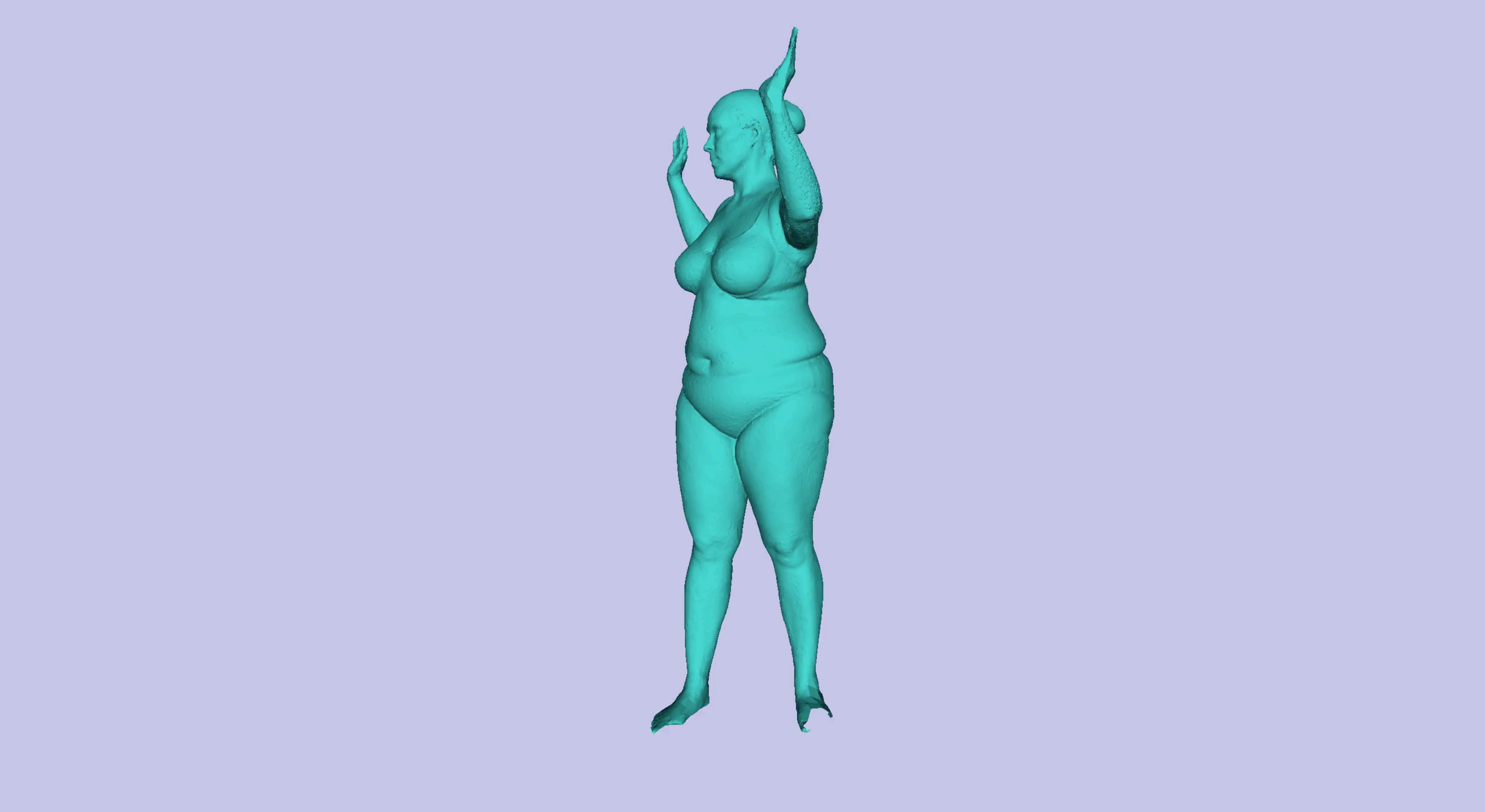


In Lots of Poses



To Learn a Model We Need Correspondence

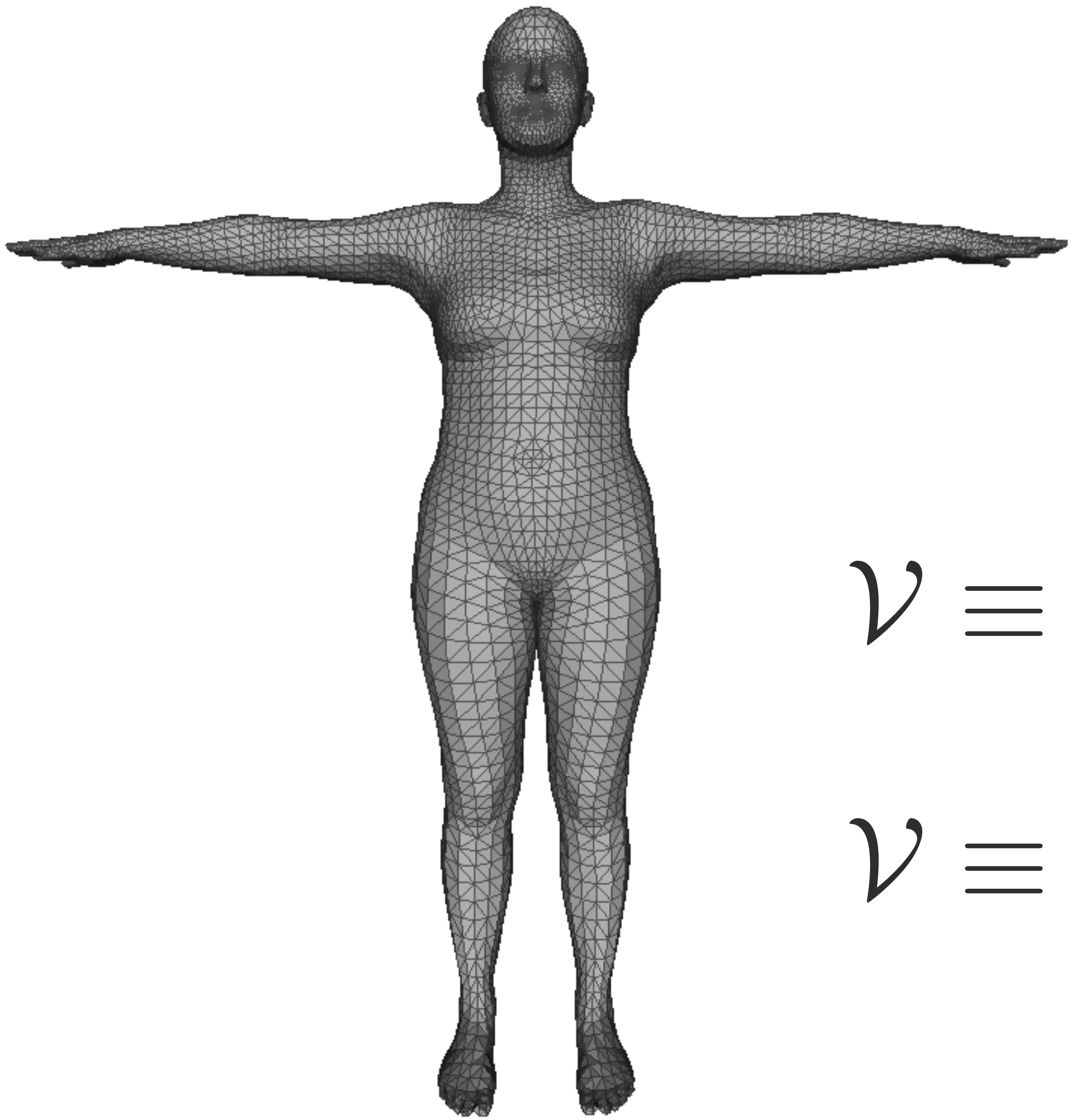




3D Body Models Basics

- Today: Geometry and alignment
  - Points, triangles and triangulated meshes
  - Similarity transformations
  - Procrustes Method

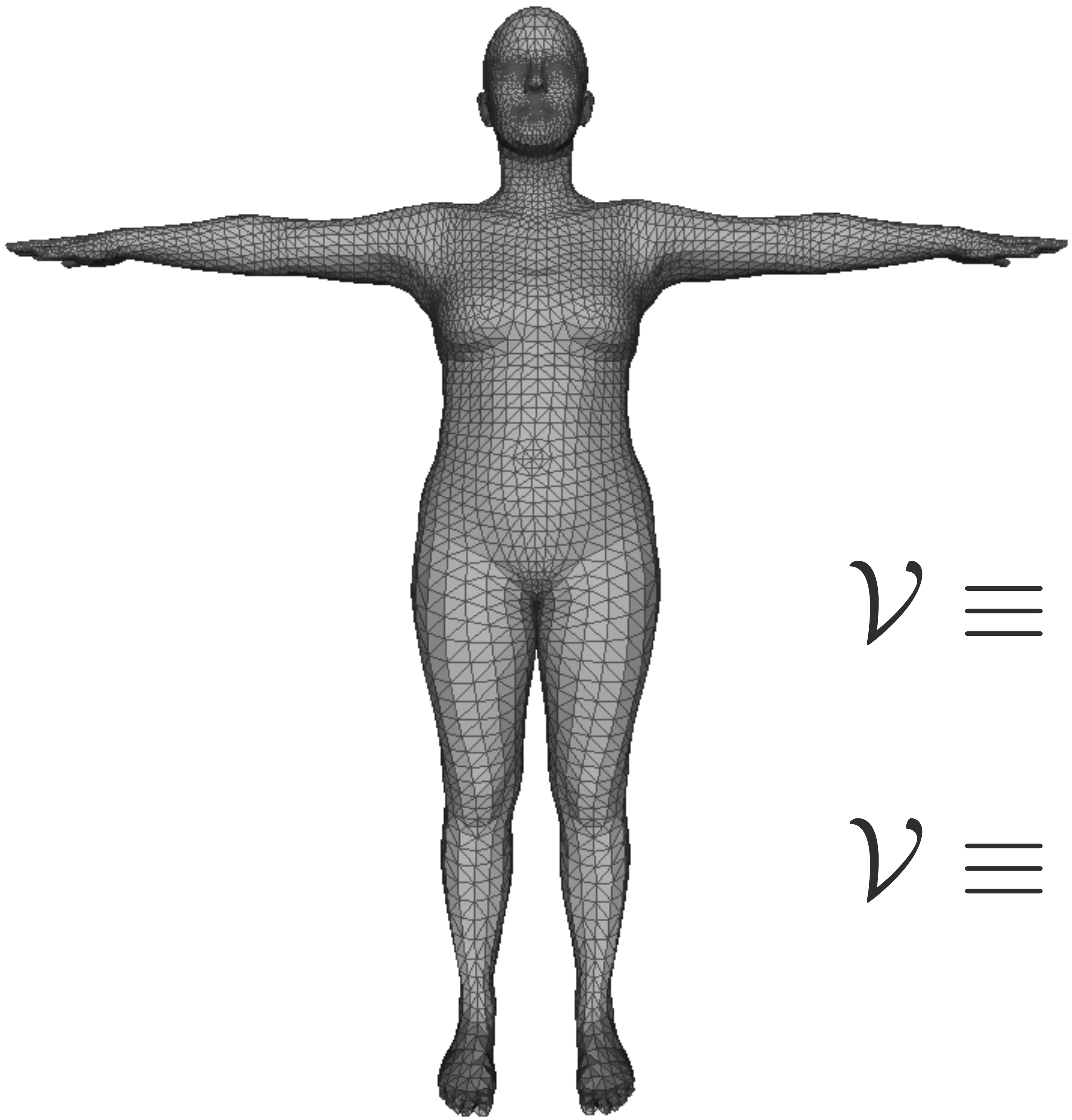
Triangulated Mesh



$$\mathcal{V} \equiv \begin{cases} \boxed{\mathbf{F} \in \mathbb{N}^{M \times 3}} \\ \mathbf{V} \in \mathbb{R}^{N \times 3} \end{cases}$$

$$\mathcal{V} \equiv \mathbb{R}^2 \to \mathbb{R}^3$$

Triangulated Mesh



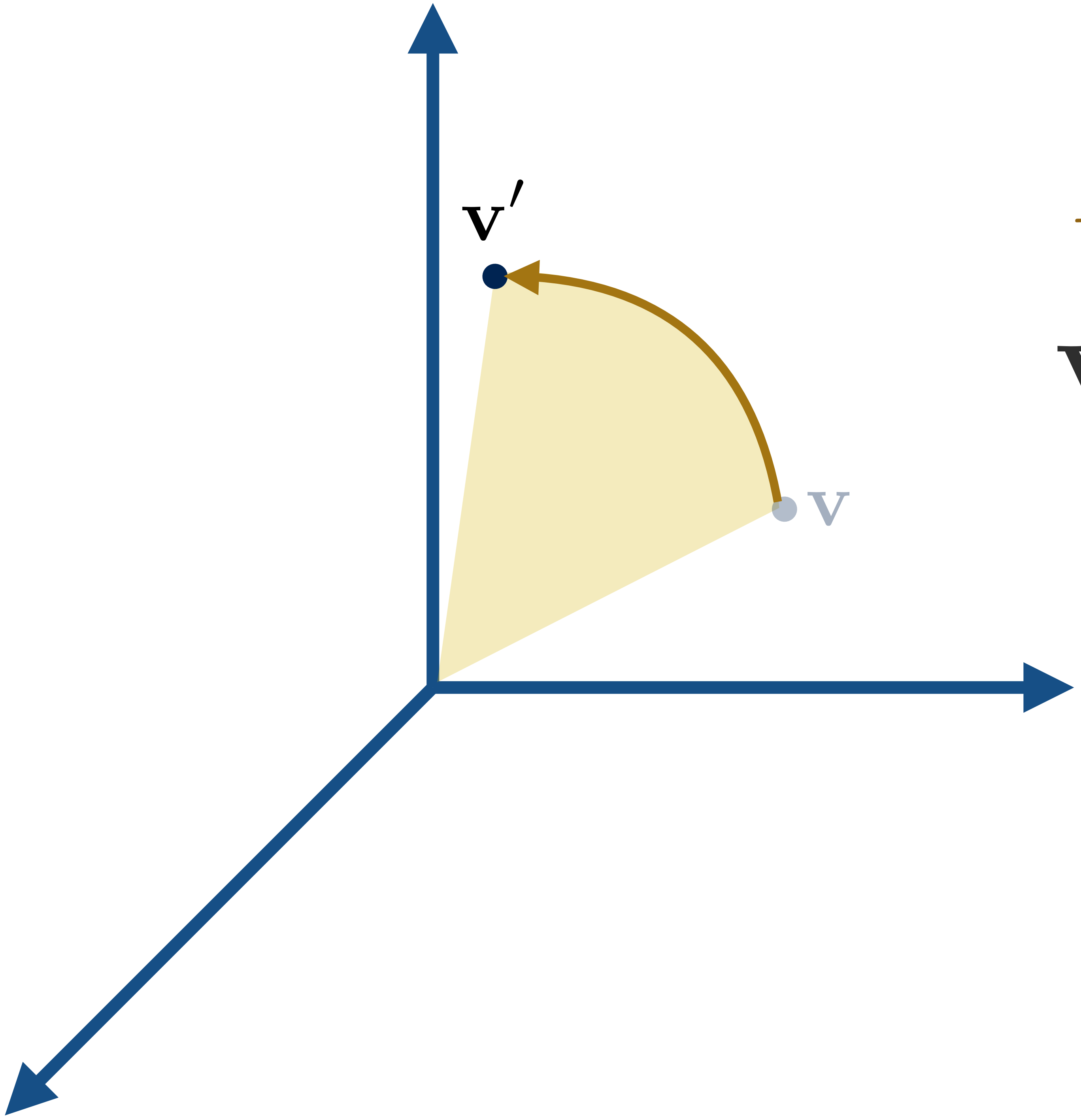
$$\mathcal{V} \equiv \begin{cases} \mathbf{F} \in \mathbb{N}^{M \times 3} \\ \mathbf{V} \in \mathbb{R}^{N \times 3} \end{cases}$$

$$\mathcal{V} \equiv \mathbb{R}^2 \rightarrow \mathbb{R}^3$$

Similarity deformations

- How can we rigidly transform a set of points ?
  - Translate it $\mathbf{V}' = \mathbf{V} + \mathbf{t}, \mathbf{t} \in \mathbb{R}^3$
  - Rotate it ${\mathbf{V}'}^{\mathrm{T}} = \mathbf{R} \cdot \mathbf{V}^{\mathrm{T}}, \mathbf{R} \in \mathbf{SO}(3)$
- Today we'll consider also
  - Scale it $\mathbf{V}' = s \cdot \mathbf{V}^{\mathrm{T}}, s \in \mathbb{R}$

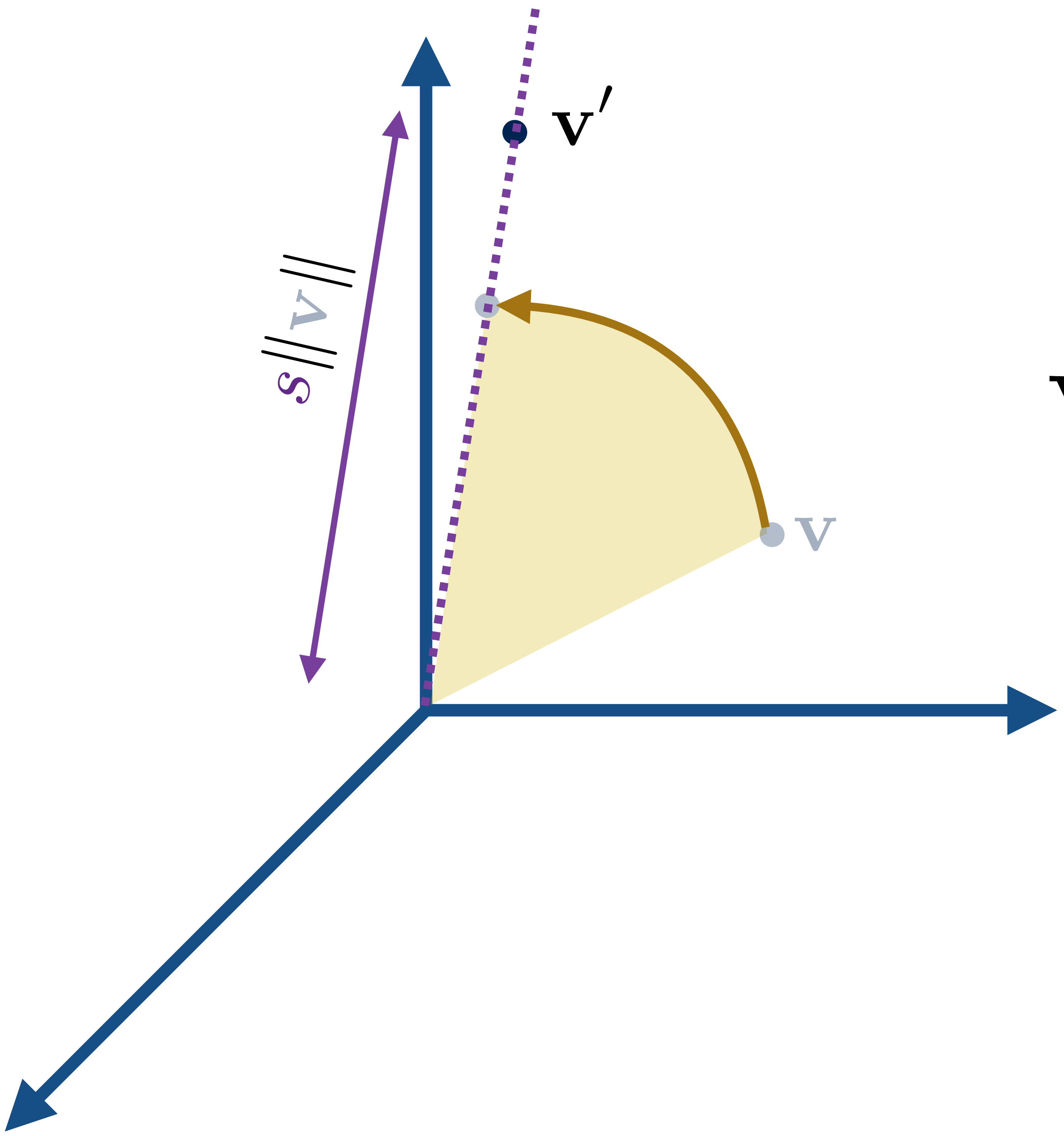
Rotation



$$R : \mathbb{R}^3 \to \mathbb{R}^3$$

$$\mathbf{v}' = R\mathbf{v}$$

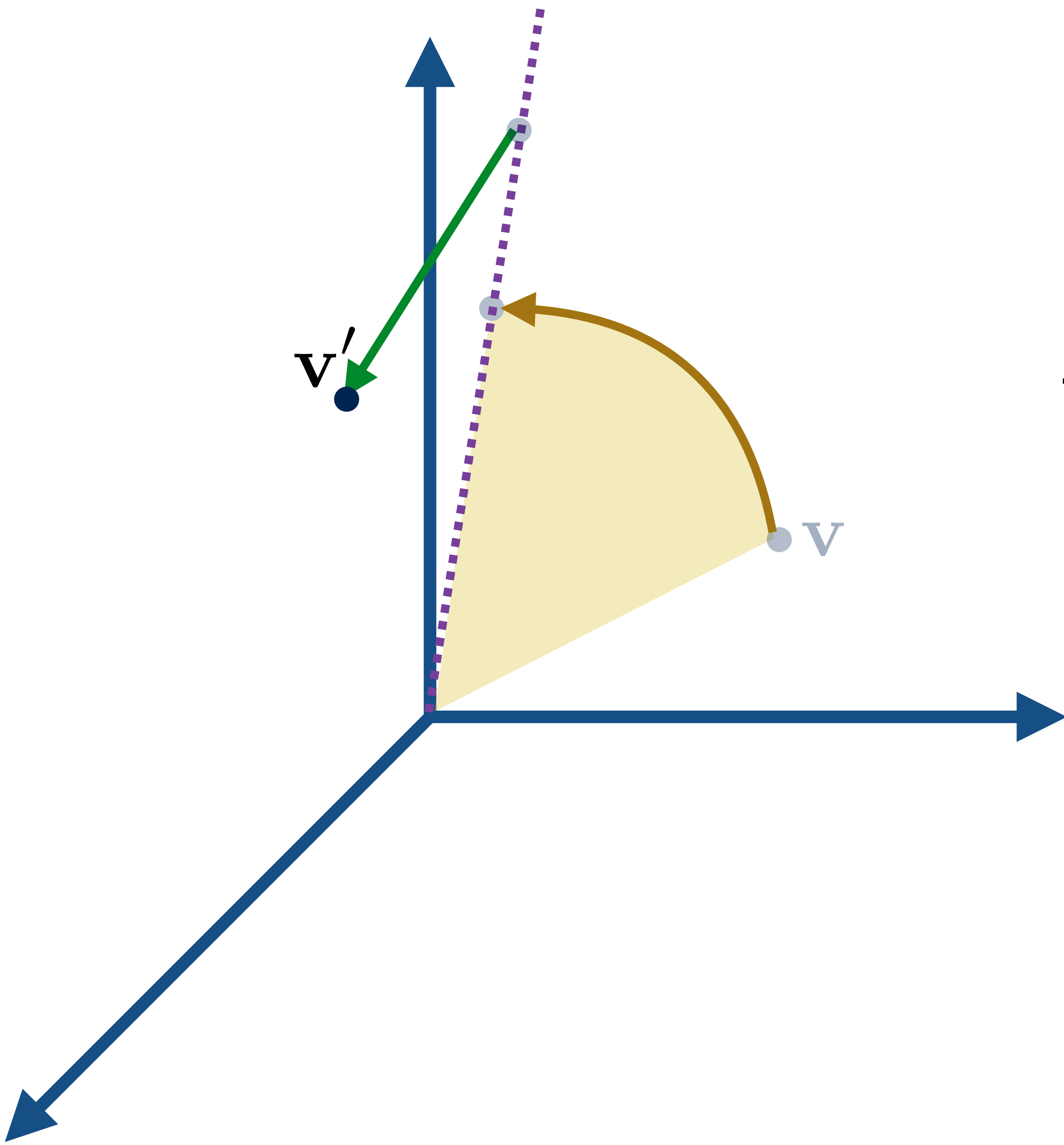
Rotation + Scale



$$R : \mathbb{R}^3 \to \mathbb{R}^3$$

$$\mathbf{v}' = sR\mathbf{v}$$

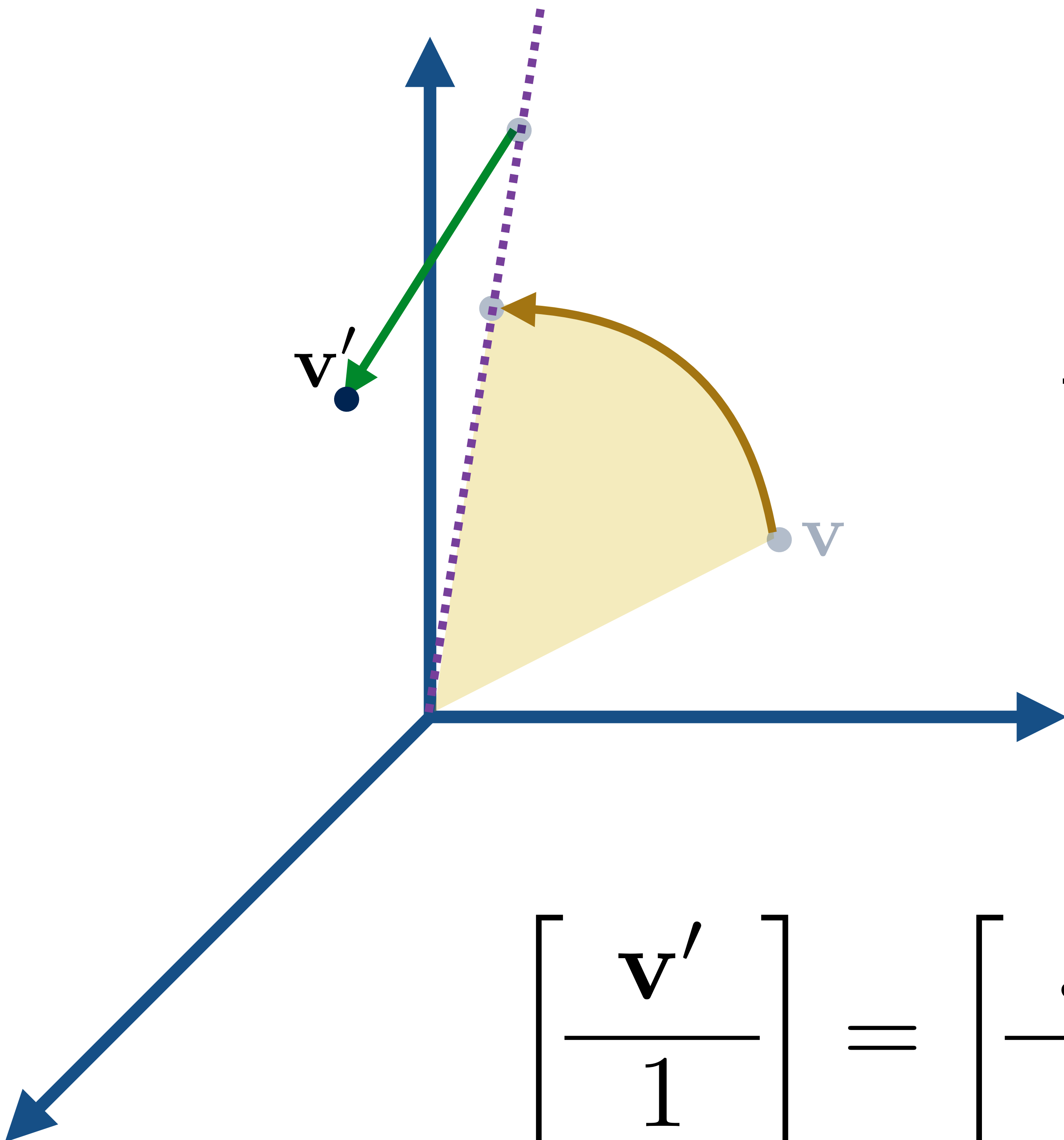
Rotation + Scale + Translation



$$R : \mathbb{R}^3 \to \mathbb{R}^3$$

$$\mathbf{v}' = sR\mathbf{v} + \mathbf{t}$$

Rotation + Scale + Translation



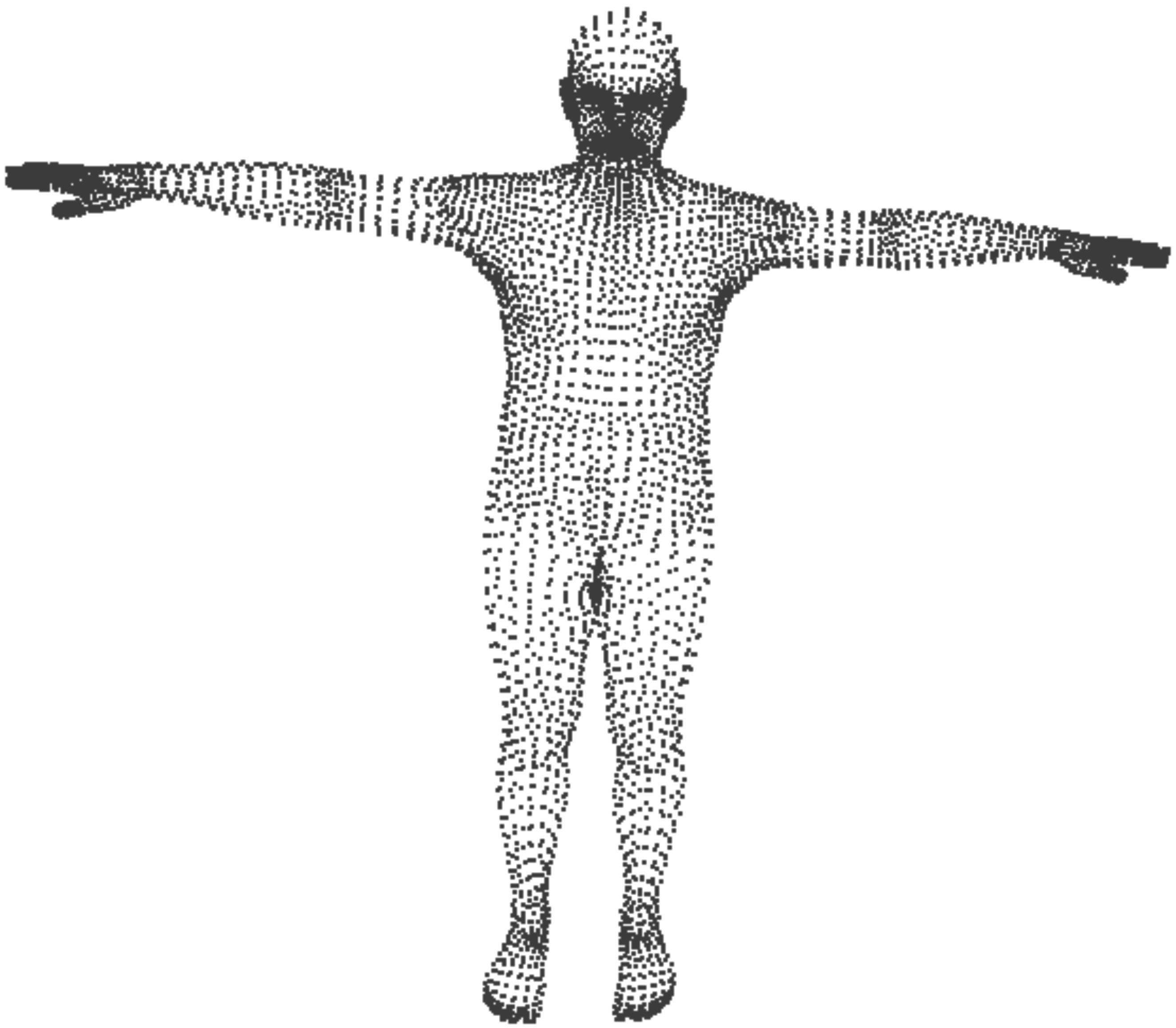
$$R : \mathbb{R}^3 \to \mathbb{R}^3$$

$$\mathbf{v}' = sR\mathbf{v} + \mathbf{t}$$

$$\left[\begin{array}{c} \mathbf{v}' \\ \hline 1 \end{array}\right] = \left[\begin{array}{c|c} s\mathbf{R} & \mathbf{t} \\ \hline 0 & 1 \end{array}\right] \left[\begin{array}{c} \mathbf{v} \\ \hline 1 \end{array}\right]$$

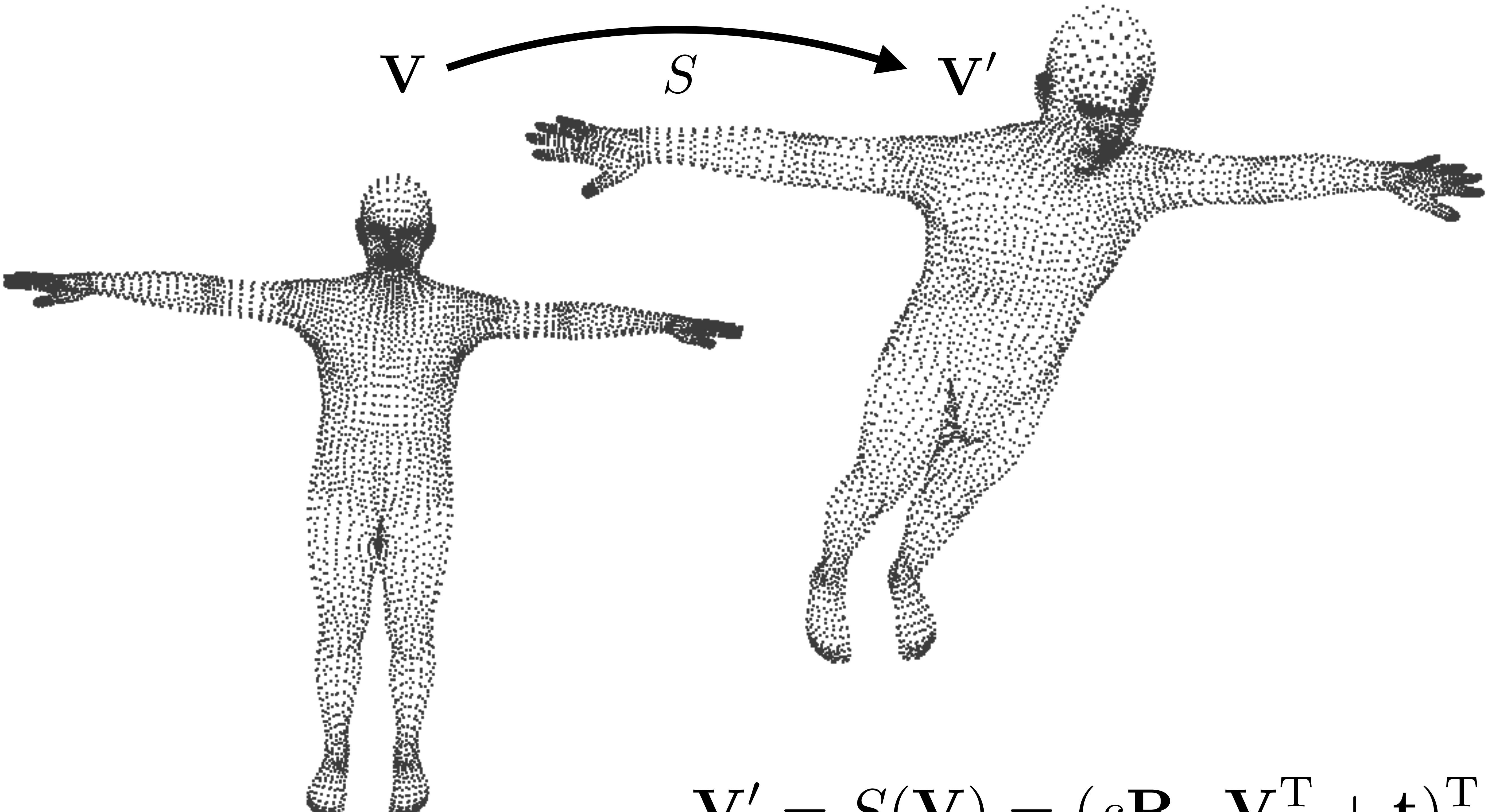
Similarity transform

$$\mathbf{v} \xrightarrow{S} \mathbf{v}'$$



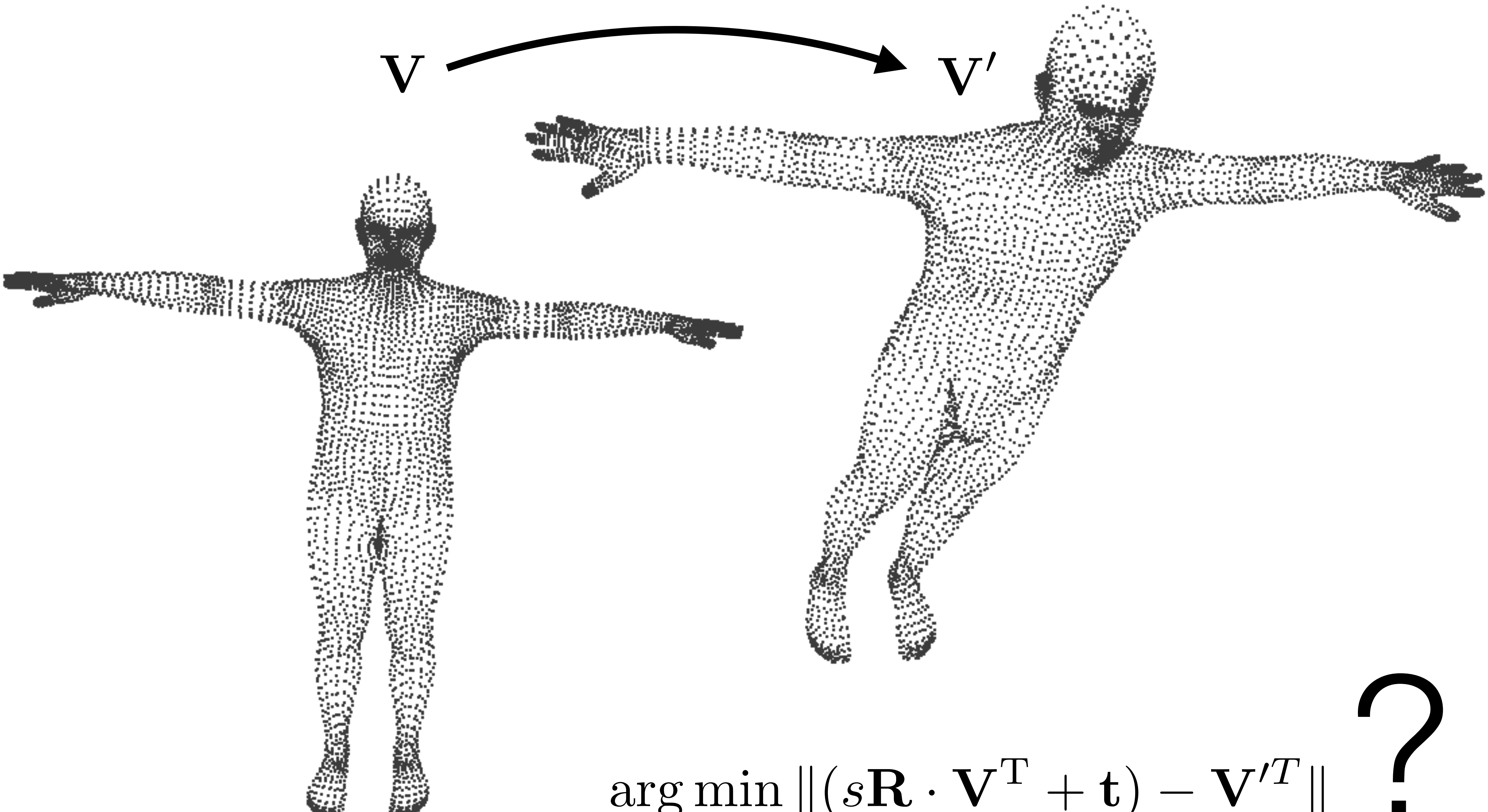
?

Similarity transform



$$\mathbf{V}' = S(\mathbf{V}) = (s\mathbf{R} \cdot \mathbf{V}^{\mathrm{T}} + \mathbf{t})^{\mathrm{T}}$$

How can we estimate S ?



$$\arg\min_{S} \|(s\mathbf{R} \cdot \mathbf{V}^{\mathrm{T}} + \mathbf{t}) - \mathbf{V}'^{T}\|$$?

Procrustes a.k.a. he who stretches

Procrustes Analysis

$$s, \mathbf{R}, \mathbf{t} = \arg\min_{s,\mathbf{R},\mathbf{t}} \sum_i \|s\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i\|^2$$

warning: change of variable names V->X, V'->Y

Steps in Similarity Alignment

Given a set of K points: Configuration

Translation normalization: Centered Configuration
(center of mass at origin)

Scale normalization: Pre-shape
(divide by Sqrt of SSQ centered coordinates)

Rotation normalization: Shape
(rotate to alignment with ref shape)

Procrustes Analysis

$$\bar{\mathbf{X}}^T \bar{\mathbf{Y}} = \mathbf{U}\Sigma\mathbf{V}^T$$

Optimal rotation obtained by computing SVD on the point cross-covariance

$$\mathbf{R} = \mathbf{U}\mathbf{V}^T$$

$$\mathbf{t} = \sum \mathbf{R}\bar{\mathbf{x}} - \bar{\mathbf{y}}$$

$$\mathbf{s} = \frac{\mathrm{tr}(\bar{\mathbf{X}}^T \bar{\mathbf{Y}})}{\|\bar{\mathbf{X}}\|^2}$$

Take-home messages

- A human body can be represented as a graphical model with pairwise terms modelling the relation between parts
- We will study a *holistic model* with a representation learned from real people scans
- To learn a statistical 3D model from data we need *correspondence* across examples
- The optimal *similarity* transformation between two point sets can be obtained with the Procrustes algorithm

Appendix: Derivation of Procrustes Analysis

Appendix Slides: Derivation of Procrustes Analysis

$$s, \mathbf{R}, \mathbf{t} = \arg\min_{s,\mathbf{R},\mathbf{t}} \sum_i \|s\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i\|^2$$

warning: change of variable names V->X, V'->Y

SVD

in general, applied to a real matrix:

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathsf{T}}$$

$$\mathbf{A} \equiv (M \times N) \text{ real}$$

$$\mathbf{U} \equiv (M \times M) \text{ orthogonal, unit norm}$$

$$\boxed{\mathbf{V}} \equiv (N \times N) \text{ orthogonal, unit norm}$$

$$\Sigma \equiv (M \times N) \text{ diagonal}$$

warning: this is not the vertex matrix!

SVD

applied to a 3D matrix

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\mathsf{T}$$

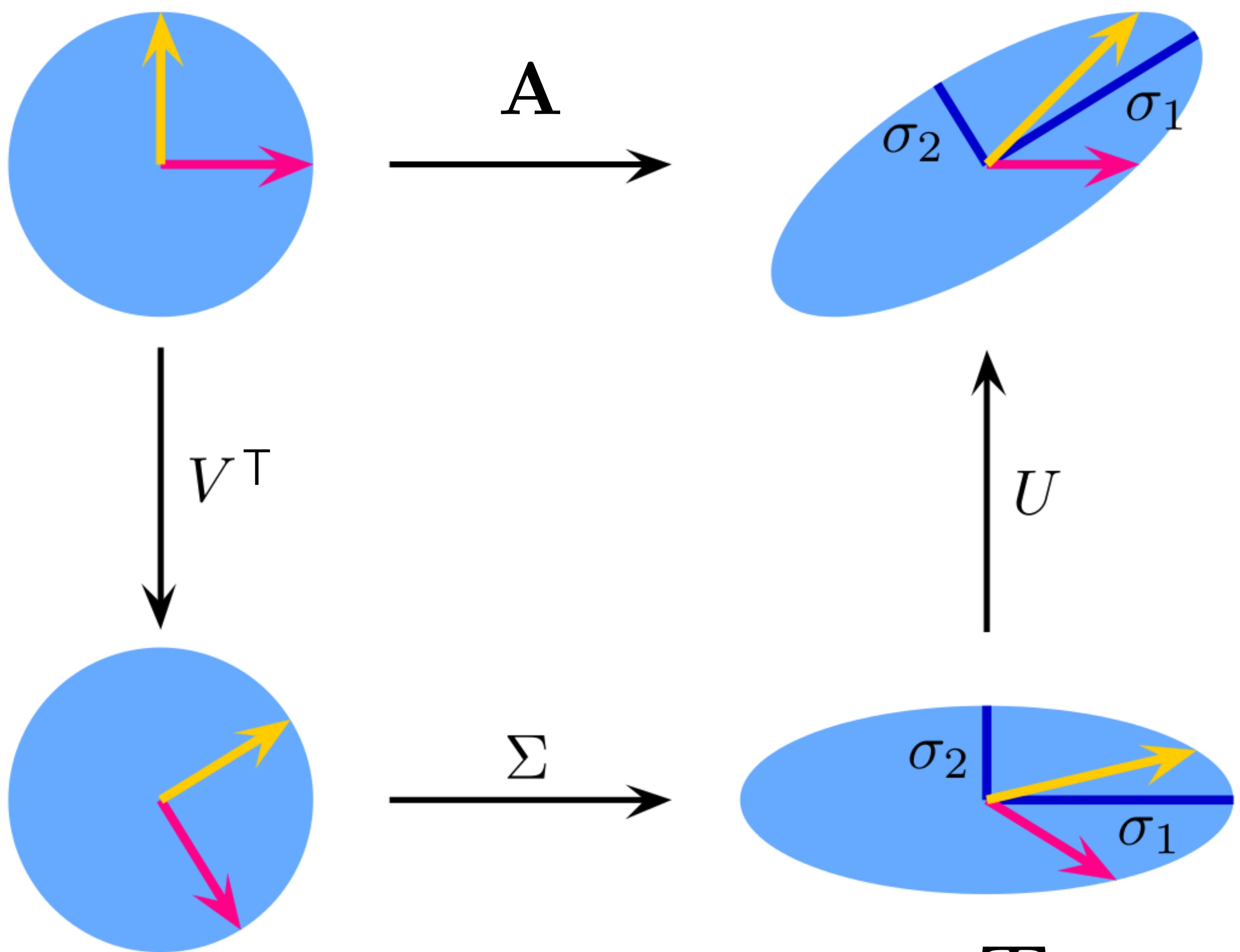
$$\mathbf{A} \equiv (3 \times 3) \text{ real}$$

$$\mathbf{U} \equiv (3 \times 3) \text{ rotation } + \text{ mirroring matrix}$$

$$\mathbf{V} \equiv (3 \times 3) \text{ rotation } + \text{ mirroring matrix}$$

$$\Sigma \equiv (3 \times 3) \text{ 3D scaling}$$

SVD



$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{W}^{\mathrm{T}}$$

Procrustes

$$\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{(N \times 3)}$$

$$\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^{(3 \times 1)}$$

$$s, \mathbf{R}, \mathbf{t} = \arg\min_{s,\mathbf{R},\mathbf{t}} E$$

$$E \equiv \sum_i \|s\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i\|^2$$

$$= \sum_i (s\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i)^\mathsf{T}(s\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i)$$

$$= \sum_i s^2\mathbf{x}_i^\mathsf{T}\mathbf{x}_i + \mathbf{t}^\mathsf{T}\mathbf{t} + \mathbf{y}_i^\mathsf{T}\mathbf{y}_i + 2s\mathbf{x}_i^\mathsf{T}\mathbf{R}^\mathsf{T}\mathbf{t} - 2s\mathbf{x}_i^\mathsf{T}\mathbf{R}^\mathsf{T}\mathbf{y}_i - 2\mathbf{t}^\mathsf{T}\mathbf{y}_i$$

Let’s call V and V’, X and Y

minimize the L2 distance between transformed source points and target points

Procrustes

If we remove the elements that do not depend on the translation and solve for t

$$\mathbf{t} = \arg\min_{\mathbf{t}} E = \arg\min_{\mathbf{t}} \sum_i \cancel{s^2 \mathbf{x}_i^\mathsf{T} \mathbf{x}_i} + \mathbf{t}^\mathsf{T}\mathbf{t} + \cancel{\mathbf{y}_i^\mathsf{T}\mathbf{y}_i} + 2s\mathbf{x}_i^\mathsf{T}\mathbf{R}^\mathsf{T}\mathbf{t} - \cancel{2s\mathbf{x}_i^\mathsf{T}\mathbf{R}^\mathsf{T}\mathbf{y}_i} - 2\mathbf{t}^\mathsf{T}\mathbf{y}_i$$

$$= \arg\min_{\mathbf{t}} \sum_i \mathbf{t}^\mathsf{T}\mathbf{t} + 2s\mathbf{x}_i^\mathsf{T}\mathbf{R}^\mathsf{T}\mathbf{t} - 2\mathbf{t}^\mathsf{T}\mathbf{y}_i$$

$$\bar{\mathbf{x}} \equiv \frac{\sum_i \mathbf{x}_i}{N}, \bar{\mathbf{y}} \equiv \frac{\sum_i \mathbf{y}_i}{N}$$ compute the centroid of the point clouds

$$\mathbf{t} = \arg\min_{\mathbf{t}} E = \arg\min_{\mathbf{t}} (\mathbf{t}^\mathsf{T}(2s\mathbf{R}\bar{\mathbf{x}} + \mathbf{t} - 2\bar{\mathbf{y}})) = \boxed{\bar{\mathbf{y}} - s\mathbf{R}\bar{\mathbf{x}}}$$

So given s and R, we can compute the translation t

Procrustes

subtract the centroid from the points to obtain a simpler expression for E

$$\bar{\mathbf{x}}_i \equiv \mathbf{x}_i - \bar{\mathbf{x}}, \bar{\mathbf{y}}_i \equiv \mathbf{y}_i - \bar{\mathbf{y}}$$

$$E = \sum_i \|s\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i\|^2 = \sum_i \|s\mathbf{R}\bar{\mathbf{x}}_i - \bar{\mathbf{y}}_i\|^2 = \sum_i s^2 \bar{\mathbf{x}}_i{}^\mathsf{T} \bar{\mathbf{x}}_i + \bar{\mathbf{y}}_i{}^\mathsf{T} \bar{\mathbf{y}}_i - 2s\bar{\mathbf{x}}_i{}^\mathsf{T} \mathbf{R}^\mathsf{T} \bar{\mathbf{y}}_i$$

$$\mathbf{A} \equiv \bar{\mathbf{X}}^\mathsf{T} \bar{\mathbf{Y}} = \sum_i \bar{\mathbf{x}}\bar{\mathbf{y}}^\mathsf{T}$$

define the cross-covariance of X and Y

$$\mathbf{R} = \arg\min_{\mathbf{R}} E = \arg\min_{\mathbf{R}} (\sum_i \cancel{s^2 \bar{\mathbf{x}}_i{}^\mathsf{T} \bar{\mathbf{x}}_i + \bar{\mathbf{y}}_i{}^\mathsf{T} \bar{\mathbf{y}}_i} - 2s\bar{\mathbf{x}}_i{}^\mathsf{T} \mathbf{R}^\mathsf{T} \bar{\mathbf{y}}_i)$$

$$= \arg\max_{\mathbf{R}} (\sum_i \bar{\mathbf{x}}_i{}^\mathsf{T} \mathbf{R}^\mathsf{T} \bar{\mathbf{y}}_i) = \arg\max_{\mathbf{R}} (\sum_i \mathrm{tr}(\bar{\mathbf{x}}_i \bar{\mathbf{y}}_i{}^\mathsf{T} \mathbf{R})) = \arg\max_{\mathbf{R}} (\mathrm{tr}(\mathbf{A}\mathbf{R}))$$

Procrustes

$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathsf{T}},$ through SVD

Decompose the cross-covariance with SVD

$\mathbf{R} = \mathbf{V}\mathbf{U}^{\mathsf{T}}$ because

$\mathbf{AR} = (\mathbf{U}\Sigma\mathbf{V}^{\mathsf{T}})(\mathbf{V}\mathbf{U}^{\mathsf{T}}) = \mathbf{U}\Sigma\mathbf{U}^{\mathsf{T}}$ symmetric positive semidefinite, so $\mathbf{R}$ maximizes $\mathrm{tr}(\mathbf{AR})$

Theorem 1: The trace of a square matrix is smaller than the sum of its eigenvalues, equal if an only if the matrix is symmetric and positive semi-definite

Theorem 2: For a square matrix $\mathbf{B}$, there is an orthogonal matrix $\mathbf{C}$ such that $\mathbf{BC}$ is symmetric and positive semi-definite. If $\mathbf{D}$ is any other orthogonal matrix, $\mathrm{tr}(\mathbf{BD}) \leq \mathrm{tr}(\mathbf{BC})$, equal if and only if $\mathbf{BD}$ is symmetric and positive semi-definite

formulation thanks to Dr. Dan Curtis

Procrustes

Optimize scale given the rotation

$$s = \arg\min_s E = \arg\min_s \left(\sum_i s^2 \bar{\mathbf{x}}_i^\top \bar{\mathbf{x}}_i + \cancel{\bar{\mathbf{y}}_i^\top \bar{\mathbf{y}}_i} - 2s\bar{\mathbf{x}}_i^\top \mathbf{R}^\top \bar{\mathbf{y}}_i\right)$$

$$= \arg\min_s \left(s^2 \sum_i (\bar{\mathbf{x}}_i^\top \bar{\mathbf{x}}_i) - 2s \sum_i (\bar{\mathbf{x}}_i^\top \mathbf{R}^\top \bar{\mathbf{y}}_i)\right)$$

$$= \arg\min_s (s^2 a - 2sb) = \frac{b}{a} = \frac{\bar{\mathbf{x}}_i^\top \mathbf{R}^\top \bar{\mathbf{y}}_i}{\bar{\mathbf{x}}_i^\top \bar{\mathbf{x}}_i}$$

$$= \frac{\mathrm{tr}(\mathbf{AR})}{\|\bar{\mathbf{X}}\|^2} = \boxed{\frac{\mathrm{tr}(\Sigma)}{\|\bar{\mathbf{X}}\|^2}}$$

Procrustes

whole algorithm

$$
\begin{aligned}
\mathbf{X}, \mathbf{Y} &\in \mathbb{R}^{(N \times 3)} \\
\mathbf{x}_i, \mathbf{y}_i &\in \mathbb{R}^{(3 \times 1)} \\
s, \mathbf{R}, \mathbf{t} &= \arg\min_{s,\mathbf{R},\mathbf{t}} E \\
E &\equiv \sum_i \|s\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i\|^2 \\
&= \sum_i (s\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i)^\top (s\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i) \\
&= \sum_i s^2 \mathbf{x}_i^\top \mathbf{x}_i + \mathbf{t}^\top \mathbf{t} + \mathbf{y}_i^\top \mathbf{y}_i + 2s\mathbf{x}_i^\top \mathbf{R}^\top \mathbf{t} - 2s\mathbf{x}_i^\top \mathbf{R}^\top \mathbf{y}_i - 2\mathbf{t}^\top \mathbf{y}_i \\
\mathbf{t} &= \arg\min_{\mathbf{t}} E = \arg\min_{\mathbf{t}} \sum_i \cancel{s^2 \mathbf{x}_i^\top \mathbf{x}_i} + \mathbf{t}^\top \mathbf{t} + \cancel{\mathbf{y}_i^\top \mathbf{y}_i} + 2s\mathbf{x}_i^\top \mathbf{R}^\top \mathbf{t} - \cancel{2s\mathbf{x}_i^\top \mathbf{R}^\top \mathbf{y}_i} - 2\mathbf{t}^\top \mathbf{y}_i \\
&= \arg\min_{\mathbf{t}} \sum_i \mathbf{t}^\top \mathbf{t} + 2s\mathbf{x}_i^\top \mathbf{R}^\top \mathbf{t} - 2\mathbf{t}^\top \mathbf{y}_i \\
\bar{\mathbf{x}} &\equiv \frac{\sum_i \mathbf{x}_i}{N}, \bar{\mathbf{y}} \equiv \frac{\sum_i \mathbf{y}_i}{N} \\
\mathbf{t} &= \arg\min_{\mathbf{t}} E = \arg\min_{\mathbf{t}} (\mathbf{t}^\top (2s\mathbf{R}\bar{\mathbf{x}} + \mathbf{t} - 2\bar{\mathbf{y}})) = \bar{\mathbf{y}} - s\mathbf{R}\bar{\mathbf{x}} \\
\bar{\mathbf{x}}_i &\equiv \mathbf{x}_i - \bar{\mathbf{x}}, \bar{\mathbf{y}}_i \equiv \mathbf{y}_i - \bar{\mathbf{y}} \\
E &= \sum_i \|s\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i\|^2 = \sum_i \|s\mathbf{R}\bar{\mathbf{x}}_i - \bar{\mathbf{y}}_i\|^2 = \sum_i s^2 \bar{\mathbf{x}}_i{}^\top \bar{\mathbf{x}}_i + \bar{\mathbf{y}}_i{}^\top \bar{\mathbf{y}}_i - 2s\bar{\mathbf{x}}_i{}^\top \mathbf{R}^\top \bar{\mathbf{y}}_i \\
\mathbf{A} &\equiv \bar{\mathbf{X}}^\top \bar{\mathbf{Y}} = \sum_i \bar{\mathbf{x}}\bar{\mathbf{y}}^\top \\
\mathbf{R} &= \arg\min_{\mathbf{R}} E = \arg\min_{\mathbf{R}} (\sum_i \cancel{s^2 \bar{\mathbf{x}}_i{}^\top \bar{\mathbf{x}}_i + \bar{\mathbf{y}}_i{}^\top \bar{\mathbf{y}}_i} - 2s\bar{\mathbf{x}}_i{}^\top \mathbf{R}^\top \bar{\mathbf{y}}_i) \\
&= \arg\max_{\mathbf{R}} (\sum_i \bar{\mathbf{x}}_i{}^\top \mathbf{R}^\top \bar{\mathbf{y}}_i) = \arg\max_{\mathbf{R}} (\sum_i \mathrm{tr}(\bar{\mathbf{x}}_i \bar{\mathbf{y}}_i{}^\top \mathbf{R})) = \arg\max_{\mathbf{R}} (\mathrm{tr}(\mathbf{A}\mathbf{R})) \\
\mathbf{A} &= \mathbf{U}\Sigma\mathbf{V}^\top, \text{ through SVD} \\
\mathbf{R} &= \mathbf{V}\mathbf{U}^\top \text{ because} \\
\mathbf{A}\mathbf{R} &= (\mathbf{U}\Sigma\mathbf{V}^\top)(\mathbf{V}\mathbf{U}^\top) = \mathbf{U}\Sigma\mathbf{U}^\top \text{ symmetric positive semidefinite, so } \mathbf{R} \text{ maximizes } \mathrm{tr}(\mathbf{A}\mathbf{R}) \\
s &= \arg\min_s E = \arg\min_s (\sum_i s^2 \bar{\mathbf{x}}_i{}^\top \bar{\mathbf{x}}_i + \cancel{\bar{\mathbf{y}}_i{}^\top \bar{\mathbf{y}}_i} - 2s\bar{\mathbf{x}}_i{}^\top \mathbf{R}^\top \bar{\mathbf{y}}_i) \\
&= \arg\min_s (s^2 \sum_i (\bar{\mathbf{x}}_i{}^\top \bar{\mathbf{x}}_i) - 2s \sum_i (\bar{\mathbf{x}}_i{}^\top \mathbf{R}^\top \bar{\mathbf{y}}_i)) \\
&= \arg\min_s (s^2 a - 2sb) = \frac{b}{a} = \frac{\bar{\mathbf{x}}_i{}^\top \mathbf{R}^\top \bar{\mathbf{y}}_i}{\bar{\mathbf{x}}_i{}^\top \bar{\mathbf{x}}_i} \\
&= \frac{\mathrm{tr}(\mathbf{A}\mathbf{R})}{\|\mathbf{X}\|^2} = \frac{\mathrm{tr}(\Sigma)}{\|\mathbf{X}\|^2}
\end{aligned}
$$