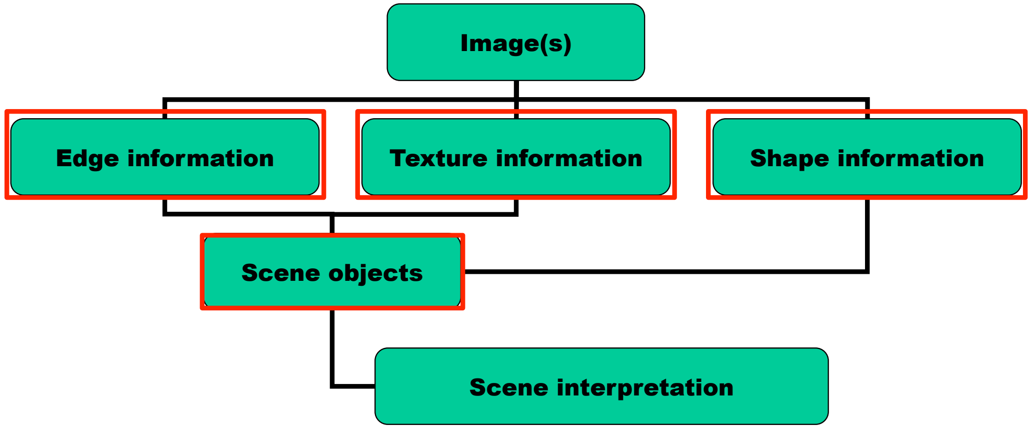
Modeling People and their Clothes

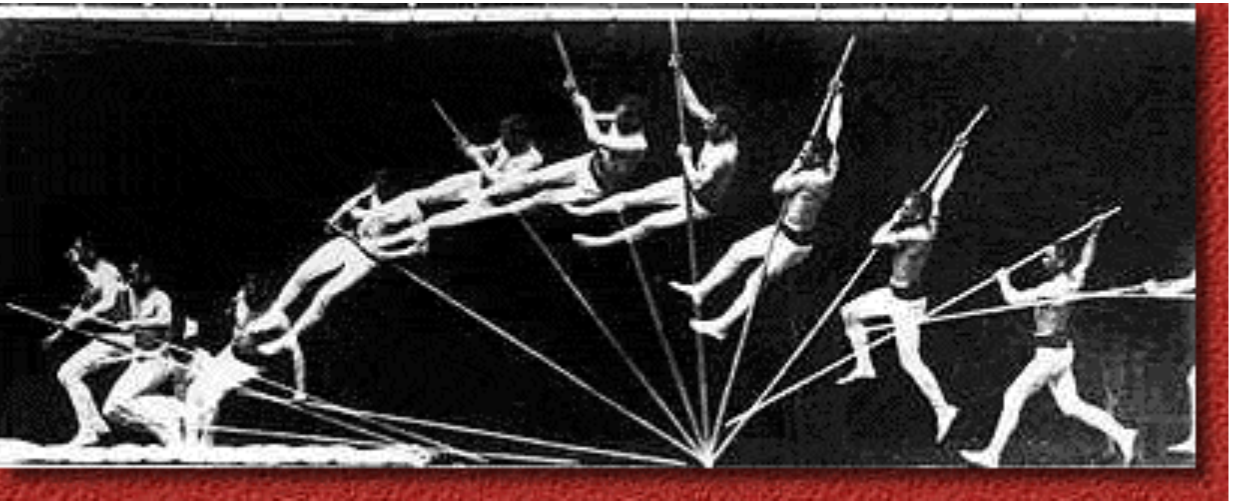
P. Fua
IC-CVLab
EPFL

A Teachable Scheme



- Capturing the body by itself.
- Modeling the clothes in relation to it.
- Handling motion and deformations.

Human Motion Analysis

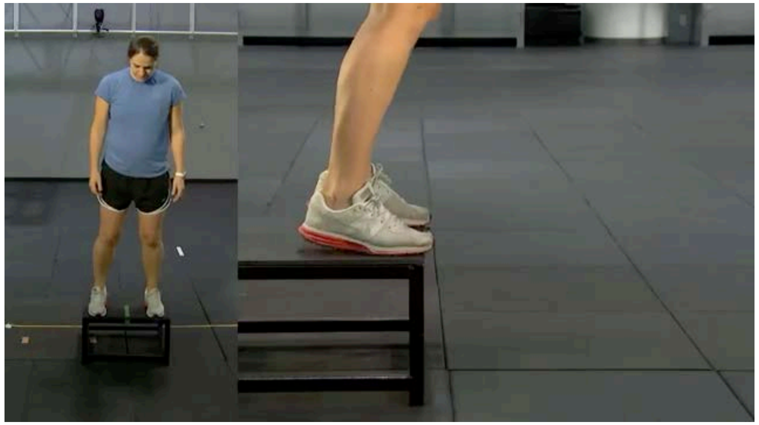


Muybridge, circa 1890

EPFL, CVLab, 2025

Applications

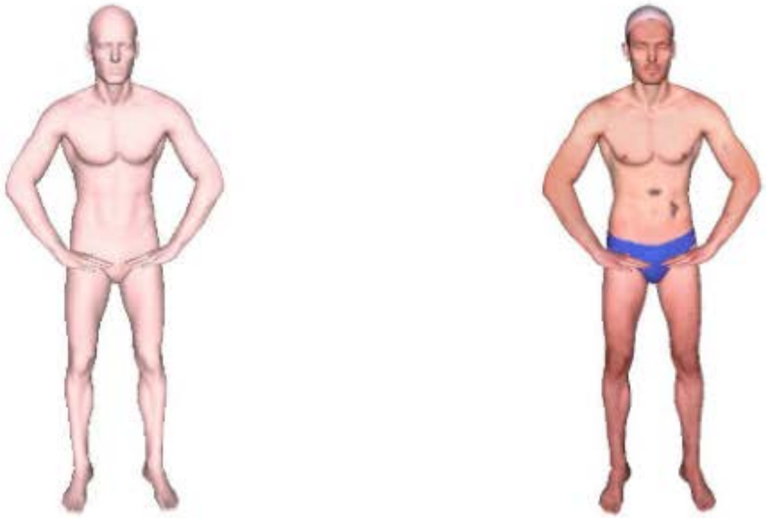
- Movies
- Fashion Design
- Sports Coaching
- Medicine
  - Enhancing performance
  - Injury prevention
  - Reeducation



Articulated Body Model

Model

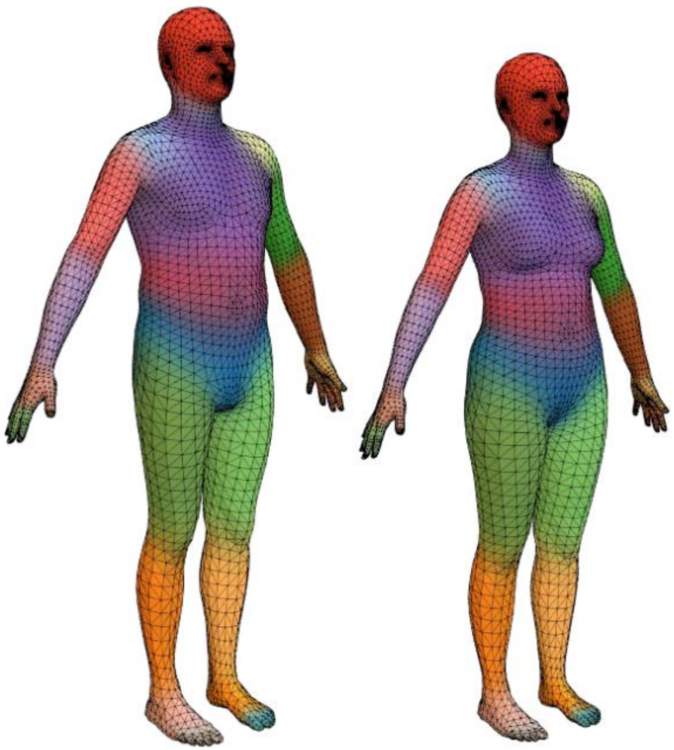
Model with texture



Pose Shape Dynamics Texture

- A model $M(\theta, \beta, \delta, A)$ takes as input a “small” number of pose, shape, and texture parameters and returns a 3D mesh.
- These parameters can be inferred from images and videos.

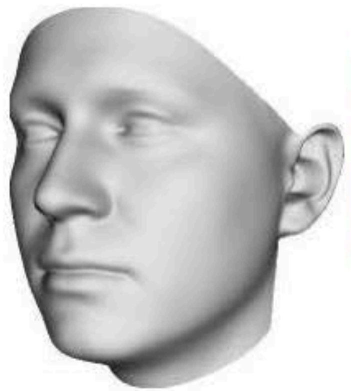
Bodies as 3D Meshes (SMPL)

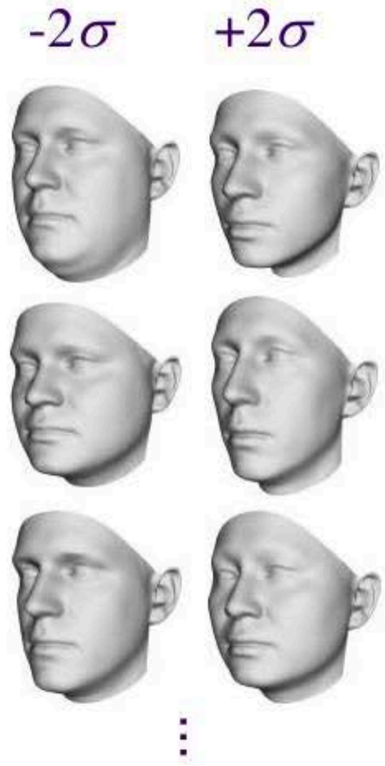


- The whole body can be represented as a low-resolution 3D mesh with 7000 vertices.
- That represents 71'000 parameters to infer from images.
- But these parameters are highly correlated.

➡The model must encode these correlations.

Reminder: PCA Face Model





$$S = \bar{S} + \sum_{i=1}^{99} \alpha_i S_j$$

$\bar{S}$: Average shape

S_i : Shape vector

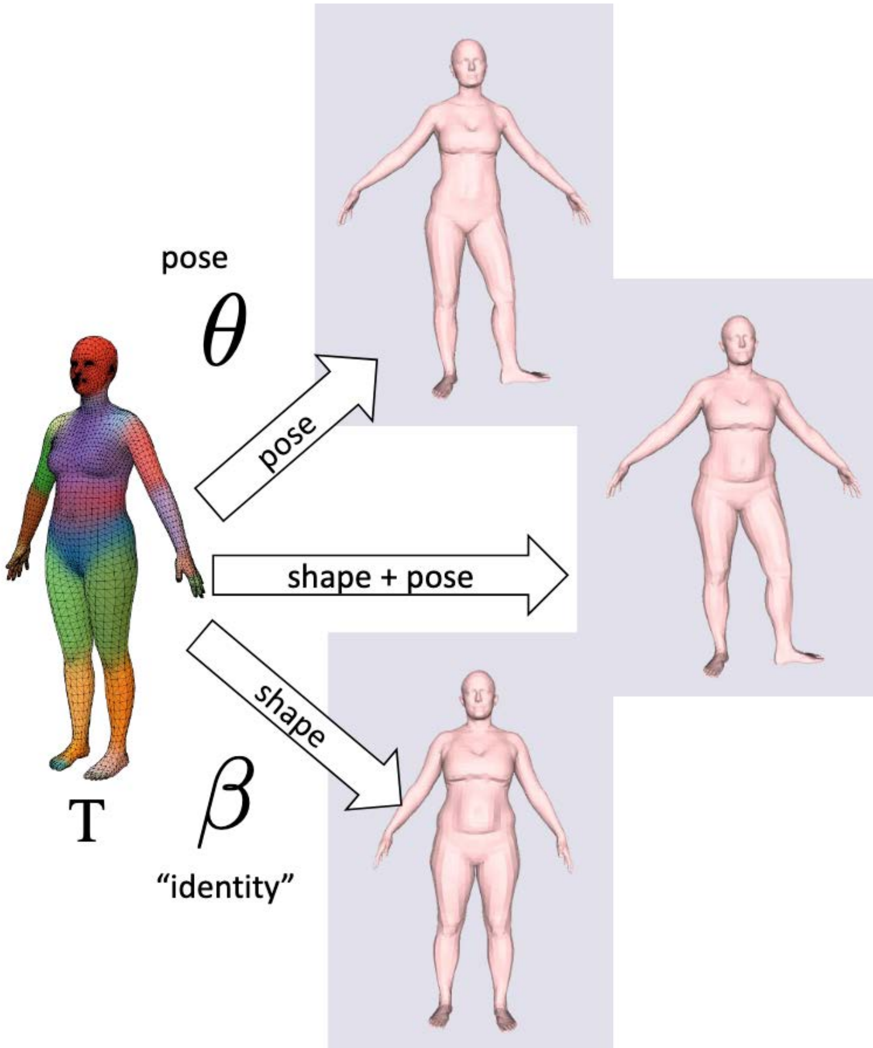
α_i : Shape coefficients

Blanz & Vetter, SIGGRAPH'99

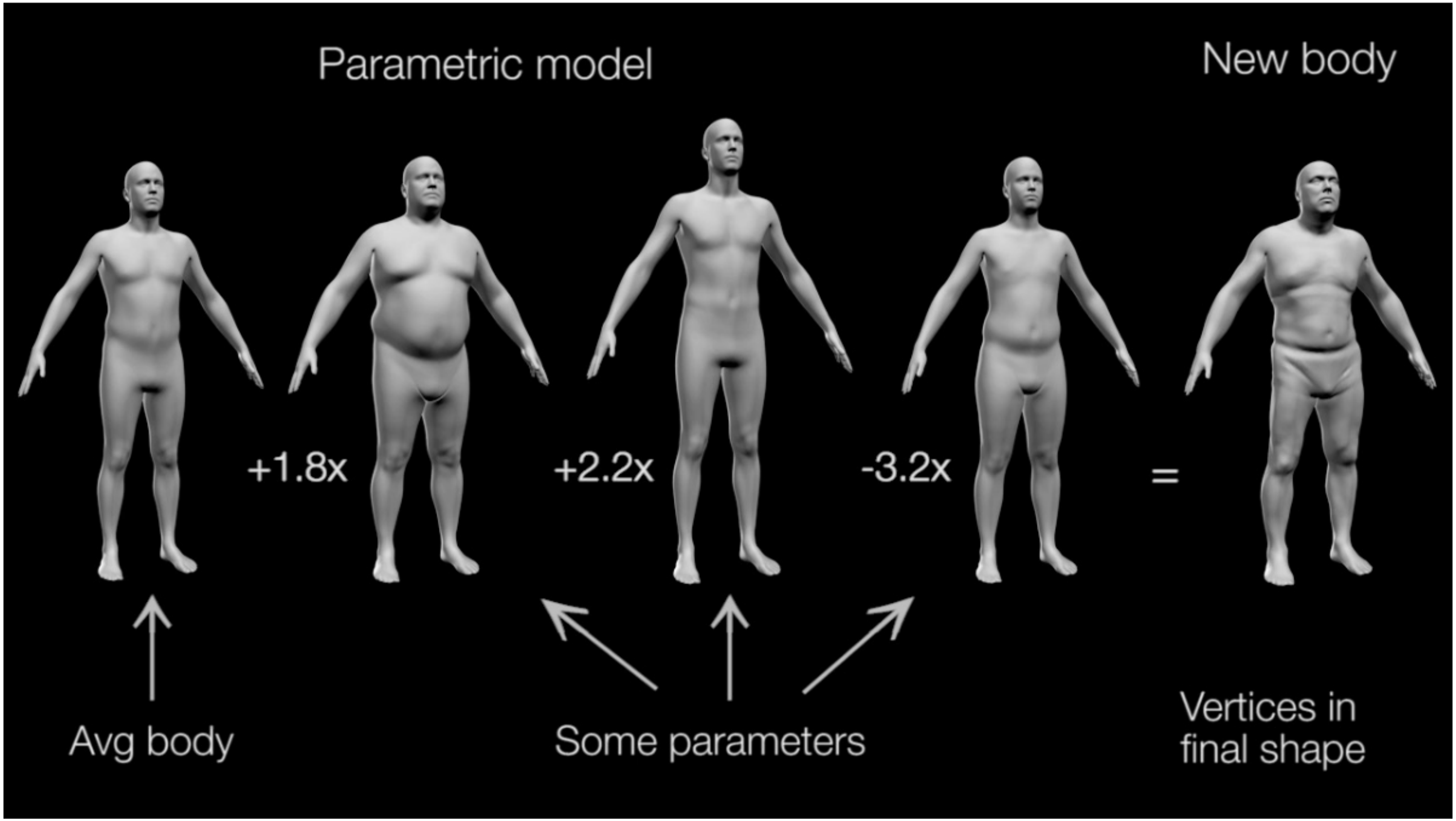
Factored Model (SCAPE)

- The model parameterizes deviations from a template mesh.
- Uses the same kind of dimensionality reduction techniques as those used to create face morphable models.
- Requires a large training database for learning purposes.

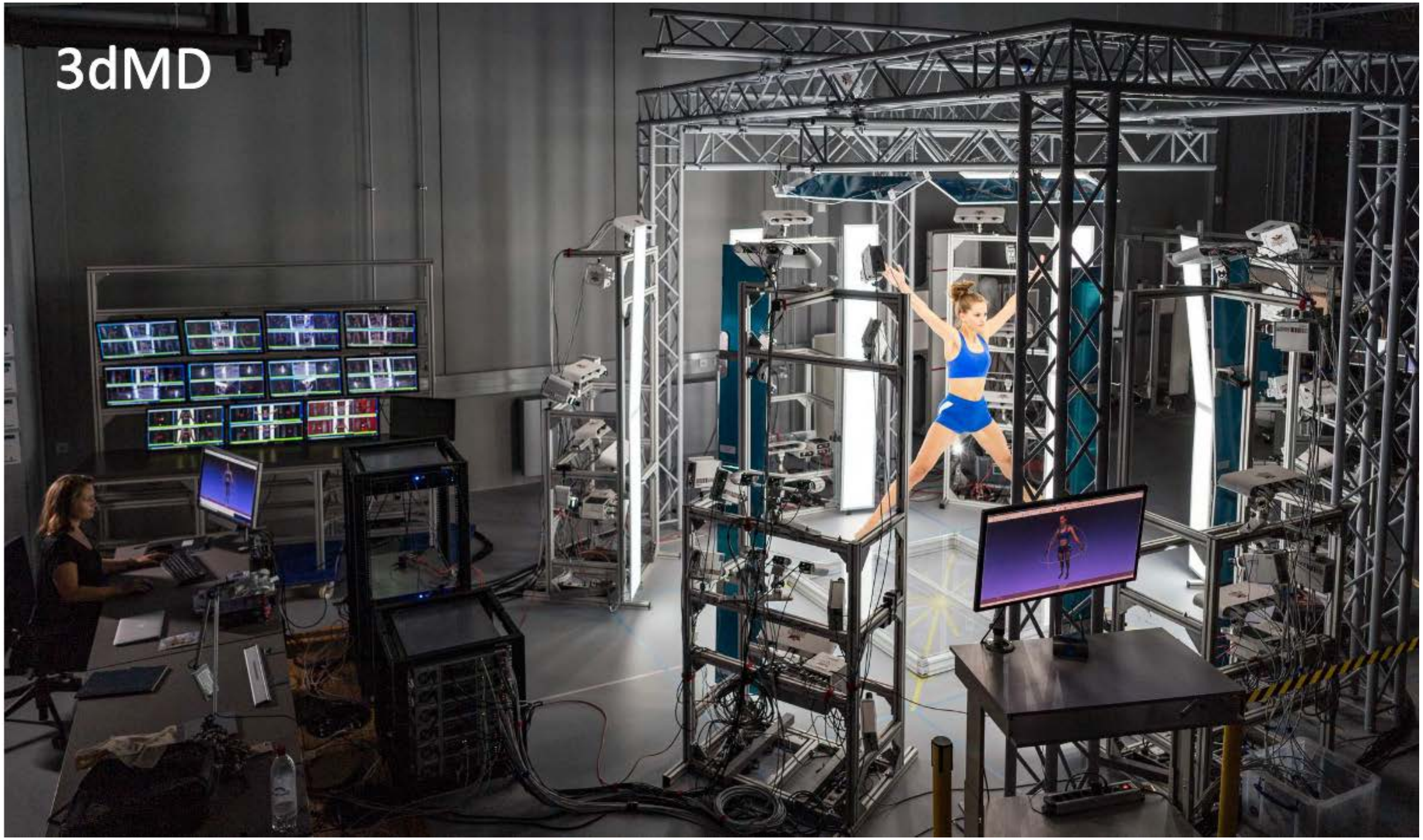
➡Simplifies learning and inference.



Changing the PCA Coefficients (SMPL)



4D Body Shapes



3dMD

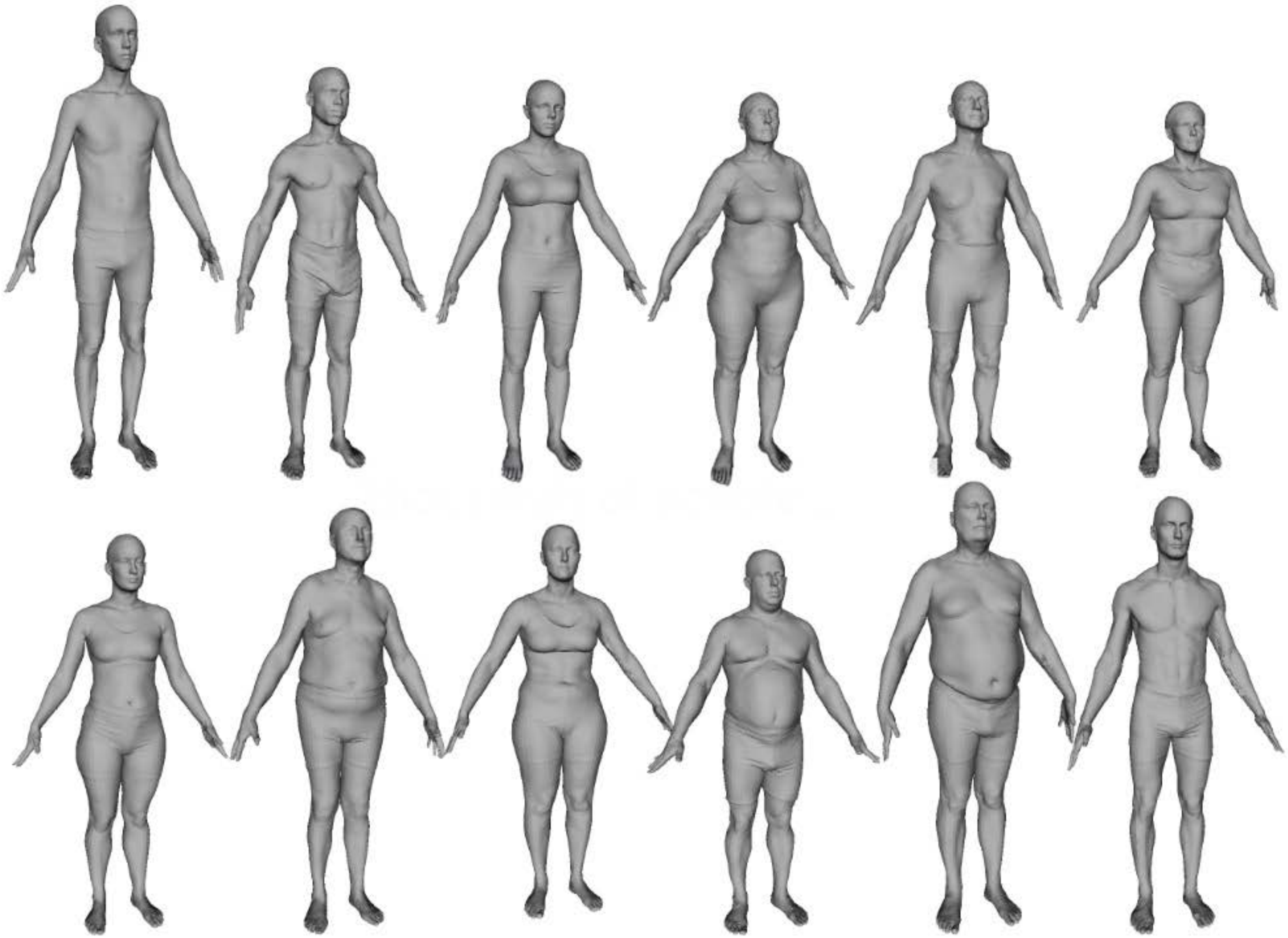
EPFL

3D at 60 FPs

University of Tübingen / MPI-Informatics:

- Thousands of people
- Thousands of poses

Many Different Body Shapes



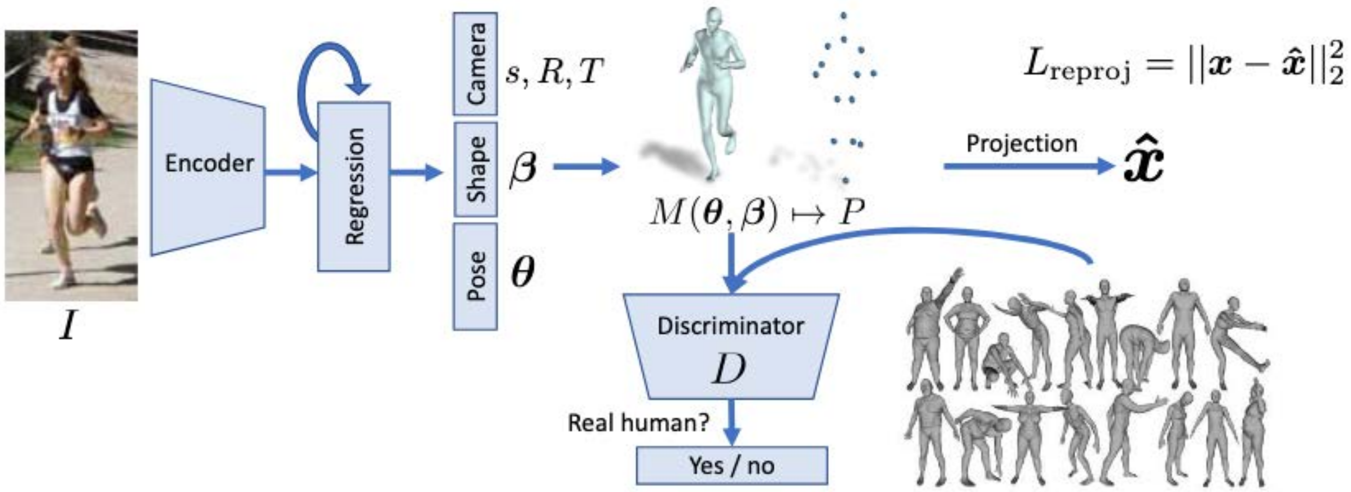
Many Different Poses

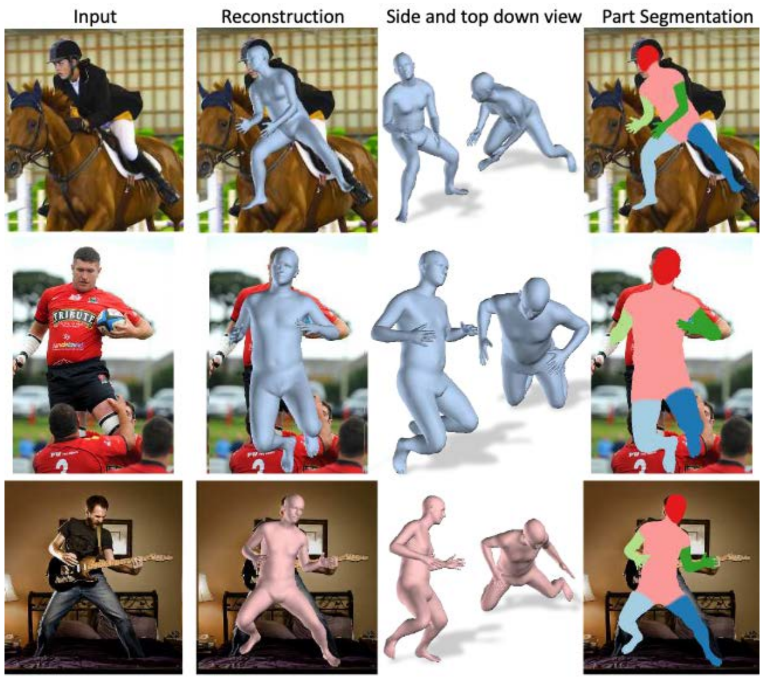
From Image to Body Pose

- Use a CNN to detect 2D joints.
- Infer SMPL parameters from those.

Not all joints can be expected to be visible!

Increasing Robustness

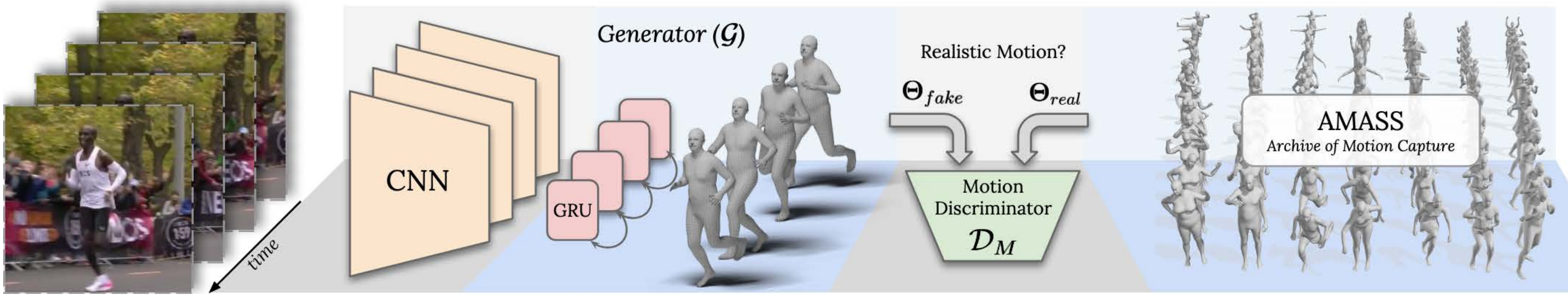




- Detect 2D joints.
- Infer SMPL parameters from those.
- Use adversarial training to ensure consistency.

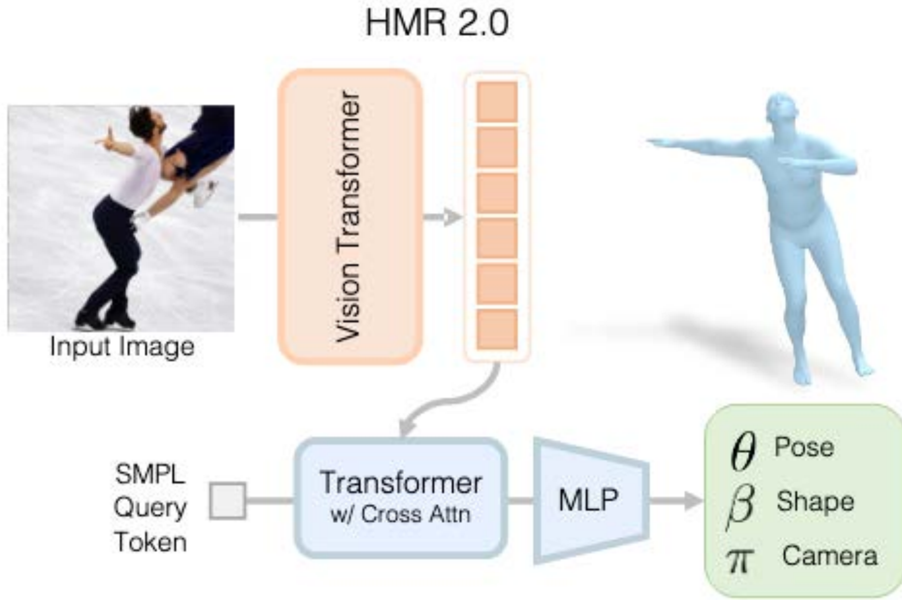
Kanazawa et al. , CVPR’18

From Video to Body Motion



- Estimate SMPL parameters from each individual video frame while enforcing temporal consistency.
- Use an adversarial network to enforce realism, given a large motion training set.

Hidden Joints



- Loose clothing can hide individual joints.
- Bring in the transformers!

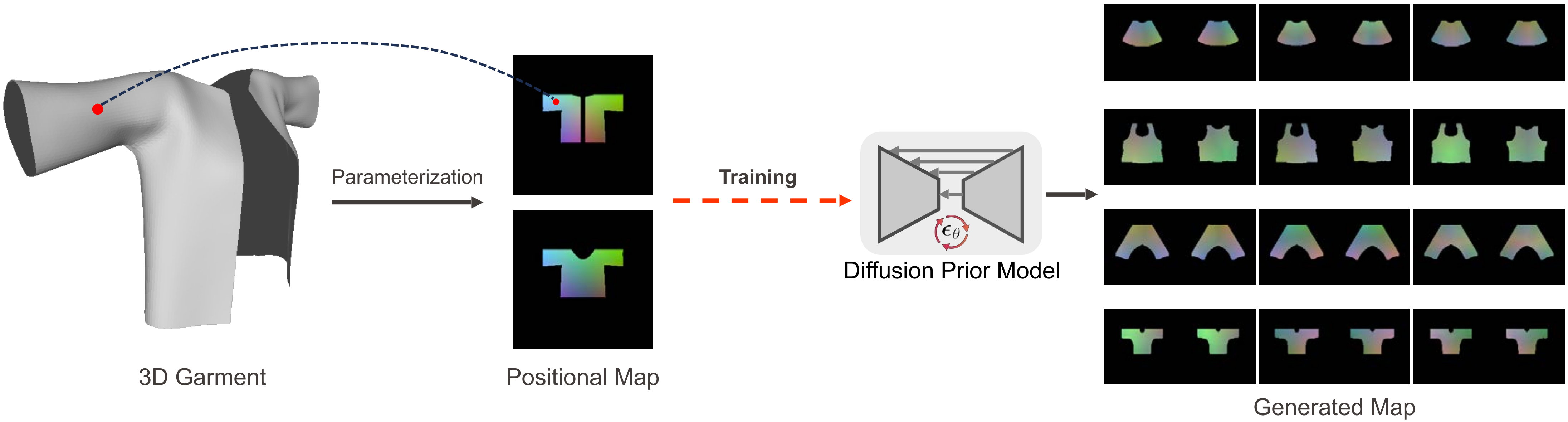
➡ Direct regression from image to SMPL parameters.

What About People Wearing Loose Clothing?



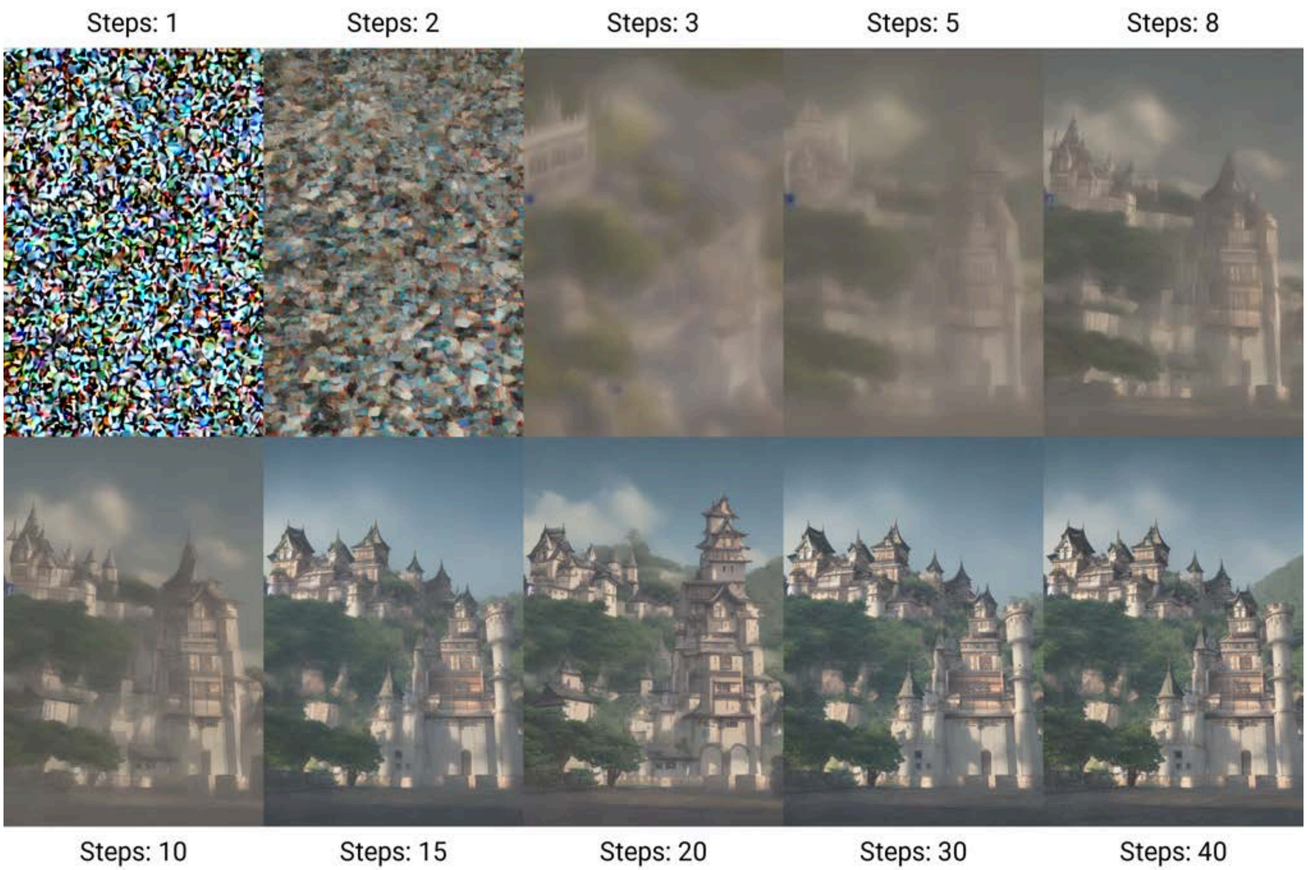

Can we also recover the shape and motion of the clothes in addition to that of the body?

Garment Parameterization



- Parameterize a 3D garment as a set of 2D positional maps.
- Train a diffusion model on it to learn the shape prior.

Diffusion / Flow Matching



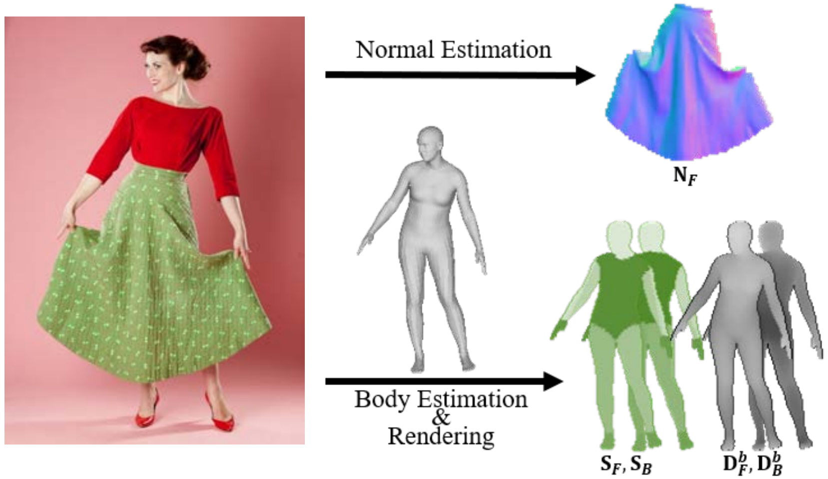
- Train a network to turn noise into a distribution that conforms to a specific prior.
- Can be "guided" to obey some constraints.

—> We use it to generate realistic clothing that matches the images.

Reconstruction Pipeline

Given an image of clothed person, its garment normal estimation $\mathbf{N}_F$ and body part/depth estimation $(\mathbf{S}_F, \mathbf{S}_B, \mathbf{D}_F^b, \mathbf{D}_B^b)$, we

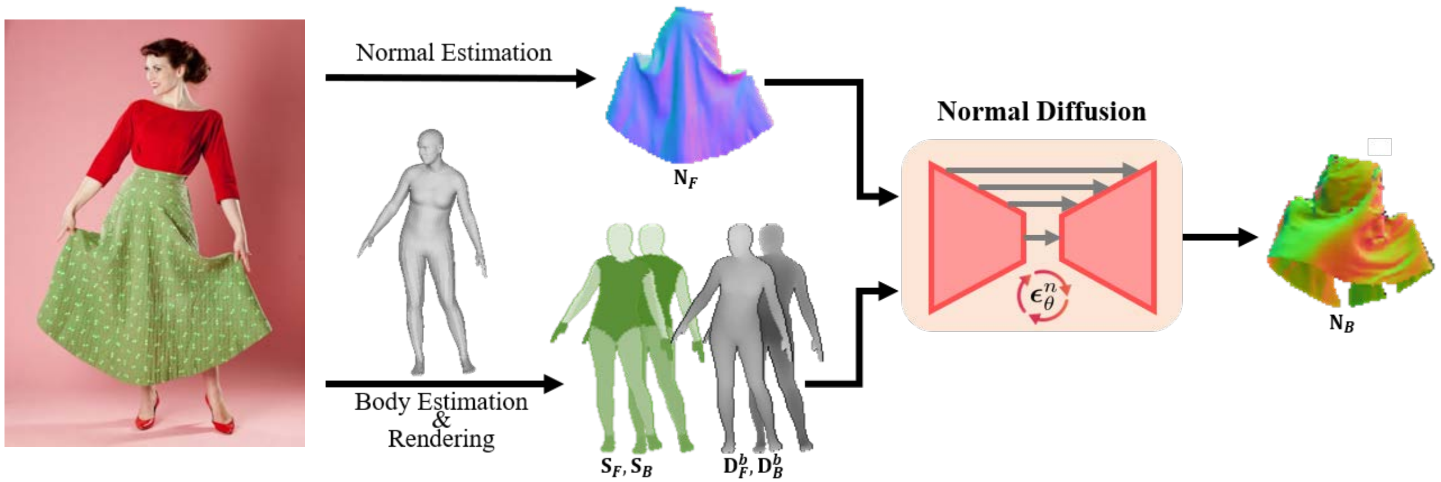
- synthesize the normal for the invisible back view $\mathbf{N}_B$
- predict the UV coordinates $(\mathbf{C}_F, \mathbf{C}_B)$ and depth $(\mathbf{D}_F^g, \mathbf{D}_B^g)$
- turn predictions to UV positional maps
- fit the prior to the positional maps for reconstruction



Reconstruction Pipeline

Given an image of clothed person, its garment normal estimation $\mathbf{N}_F$ and body part/depth estimation $(\mathbf{S}_F, \mathbf{S}_B, \mathbf{D}_F^b, \mathbf{D}_B^b)$, we

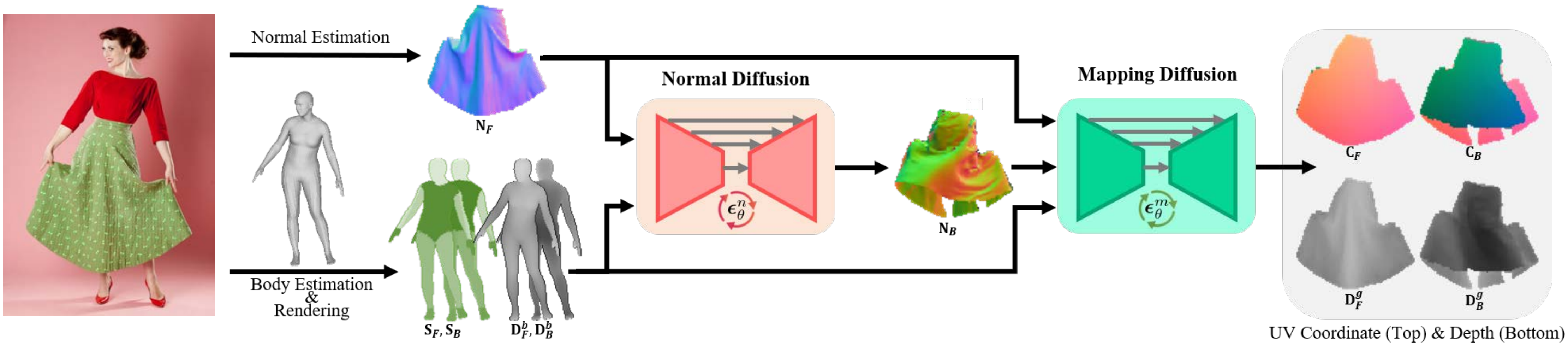
- synthesize the normal for the invisible back view $\mathbf{N}_B$
- predict the UV coordinates $(\mathbf{C}_F, \mathbf{C}_B)$ and depth $(\mathbf{D}_F^g, \mathbf{D}_B^g)$
- turn predictions to UV positional maps
- fit the prior to the positional maps for reconstruction



Reconstruction Pipeline

Given an image of clothed person, its garment normal estimation $\mathbf{N}_F$ and body part/depth estimation $(\mathbf{S}_F, \mathbf{S}_B, \mathbf{D}_F^b, \mathbf{D}_B^b)$, we

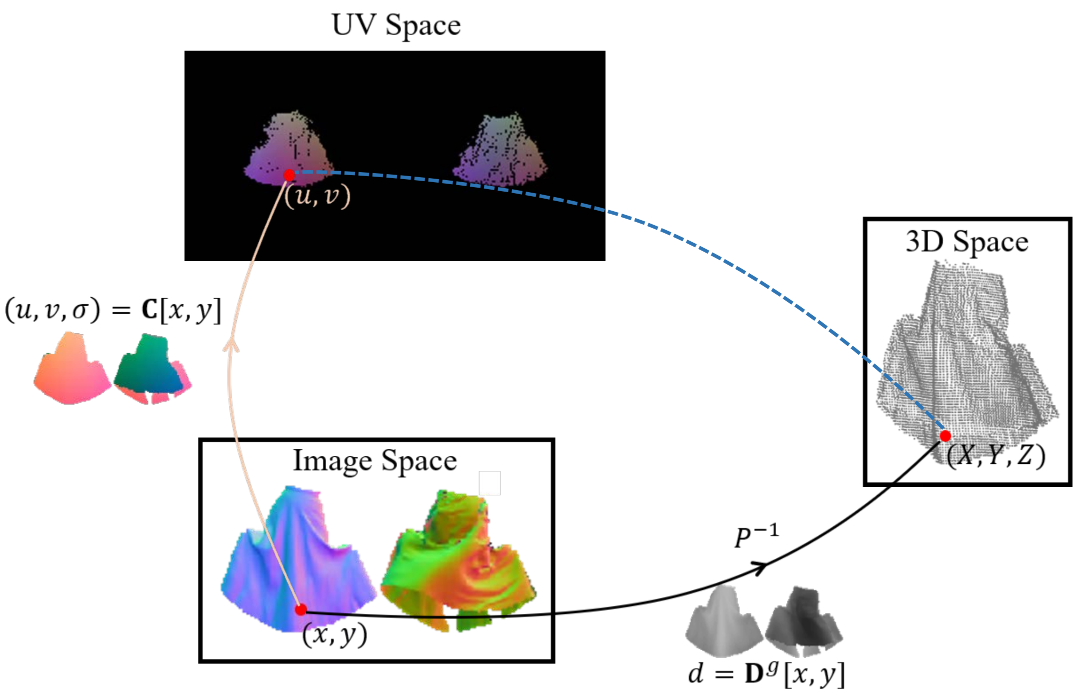
- synthesize the normal for the invisible back view N_B
- predict the UV coordinates $(\mathbf{C}_F, \mathbf{C}_B)$ and depth $(\mathbf{D}_F^g, \mathbf{D}_B^g)$
- turn predictions to UV positional maps
- fit the prior to the positional maps for reconstruction



Reconstruction Pipeline

Given an image of clothed person, its garment normal estimation $\mathbf{N}_F$ and body part/depth estimation $(\mathbf{S}_F, \mathbf{S}_B, \mathbf{D}_F^b, \mathbf{D}_B^b)$, we

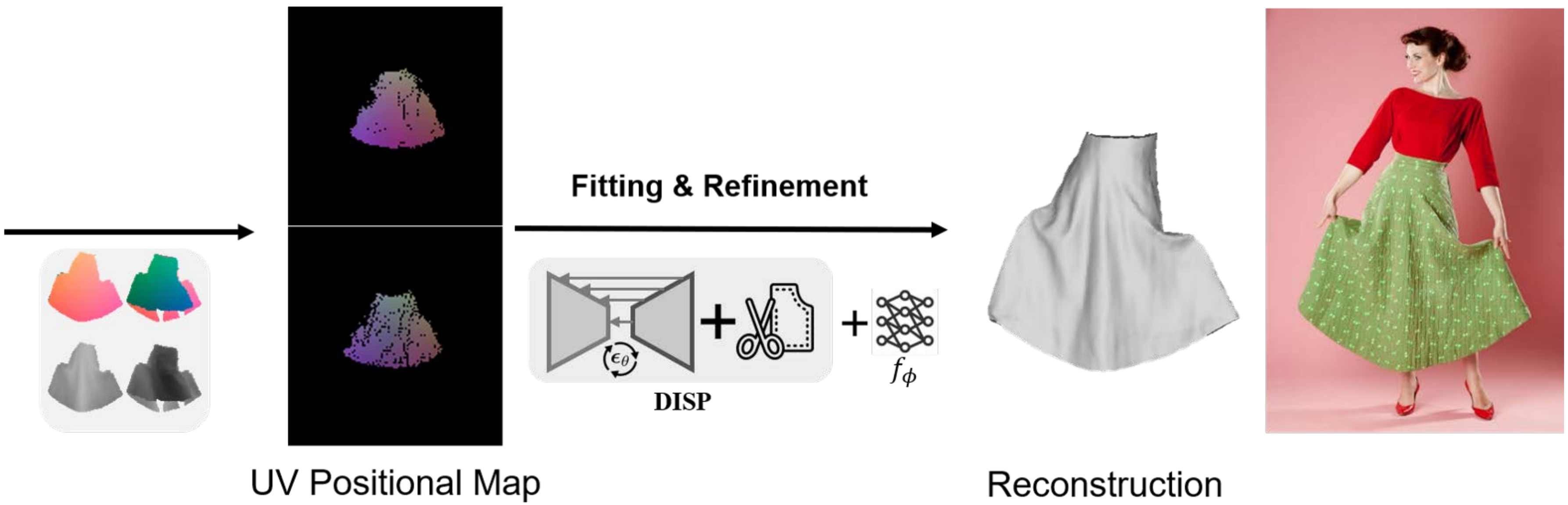
- synthesize the normal for the invisible back view $\mathbf{N}_B$
- predict the UV coordinates $(\mathbf{C}_F, \mathbf{C}_B)$ and depth $(\mathbf{D}_F^g, \mathbf{D}_B^g)$
- turn predictions to UV positional maps
- fit the prior to the positional maps for reconstruction



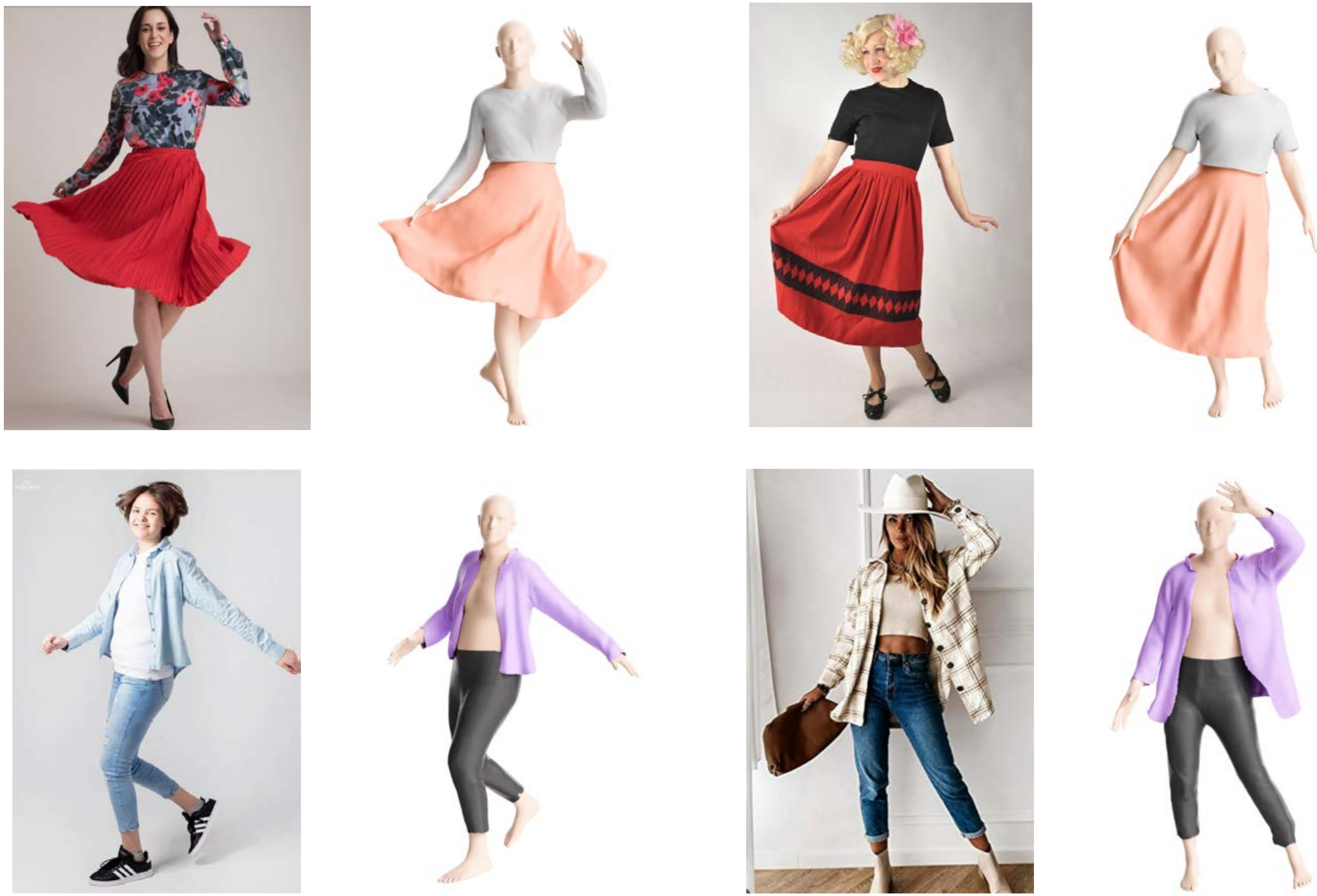
Reconstruction Pipeline

Given an image of clothed person, its garment normal estimation $\mathbf{N}_F$ and body part/depth estimation $(\mathbf{S}_F, \mathbf{S}_B, \mathbf{D}^b_F, \mathbf{D}^b_B)$, we

- synthesize the normal for the invisible back view $\mathbf{N}_B$
- predict the UV coordinates $(\mathbf{C}_F, \mathbf{C}_B)$ and depth $(\mathbf{D}^g_F, \mathbf{D}^g_B)$
- turn predictions to UV positional maps
- fit the prior to positional maps for reconstruction



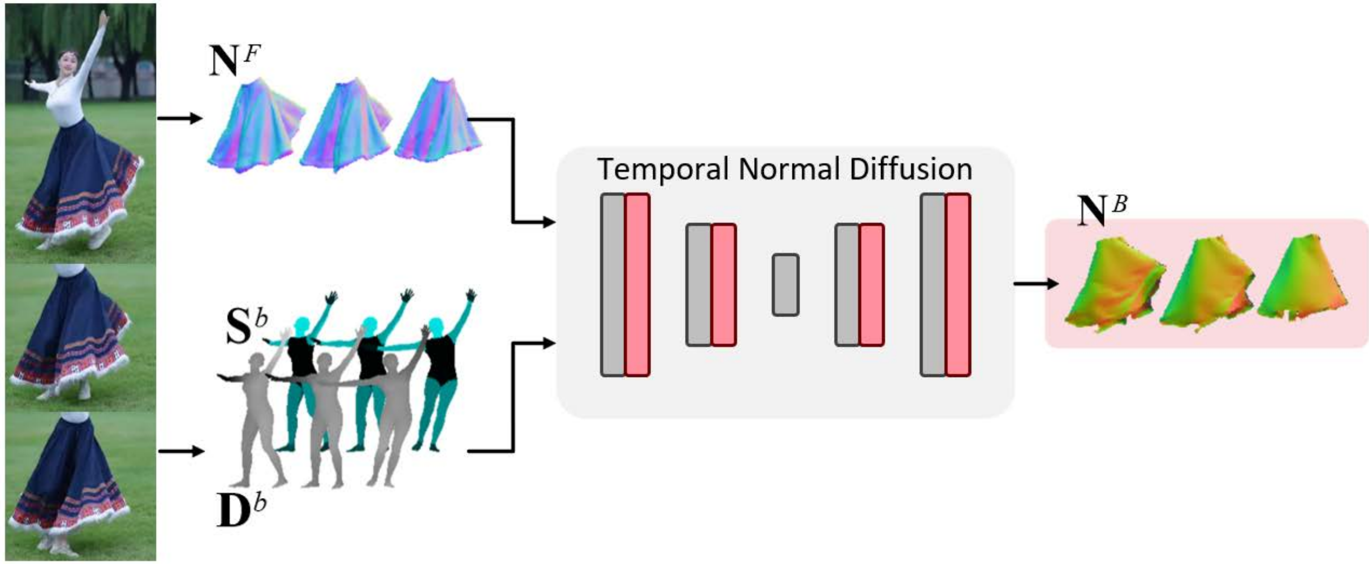
Garment Recovery from Real Images



From Images to Video

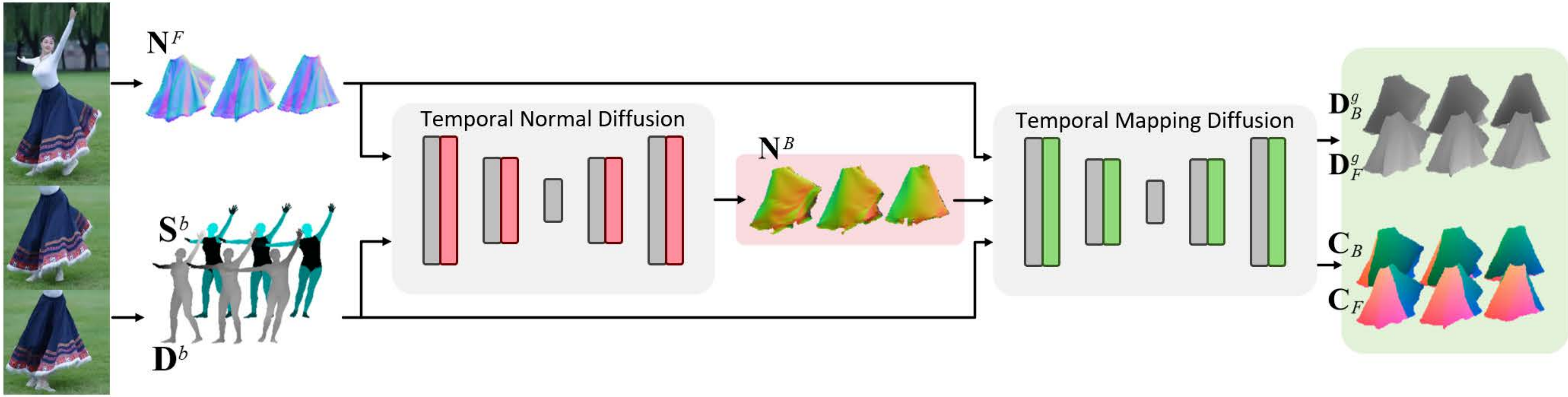
To handle video cases, we

- introduce temporal diffusion models
- enforce geometric and temporal guidance
- temporal consistency guidance, depth-to-normal guidance, interpenetration-aware guidance
- fit the prior to the positional maps with projection-based constraint for reconstruction



From Images to Videos

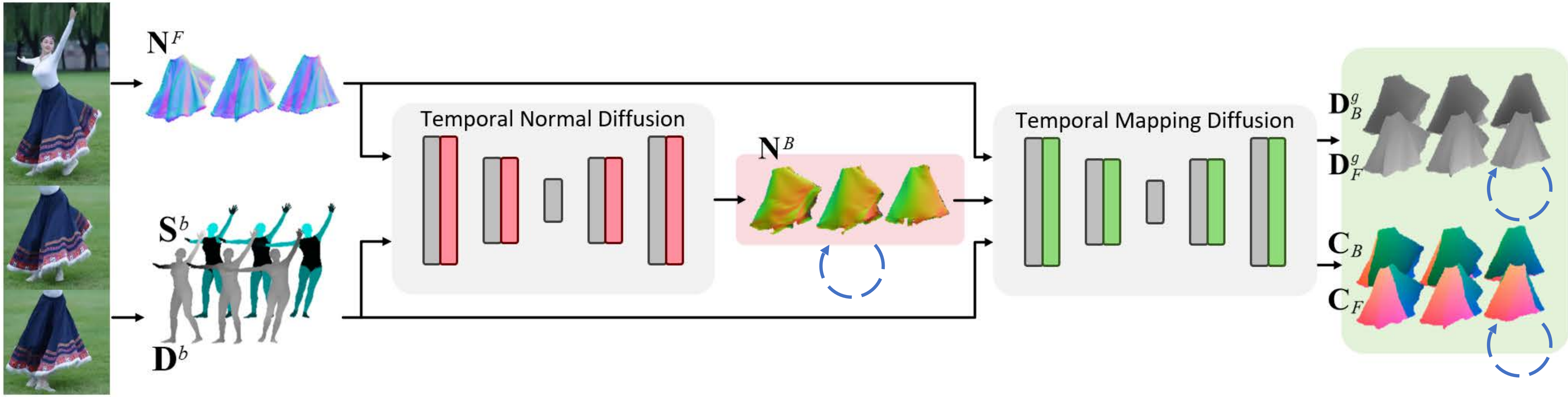
To handle video cases, we
- introduce temporal diffusion models
- enforce geometric and temporal guidance
- temporal consistency guidance, depth-to-normal guidance, interpenetration-aware guidance
- fit the prior to the positional maps with projection-based constraint for reconstruction



From Images to Videos

To handle video cases, we

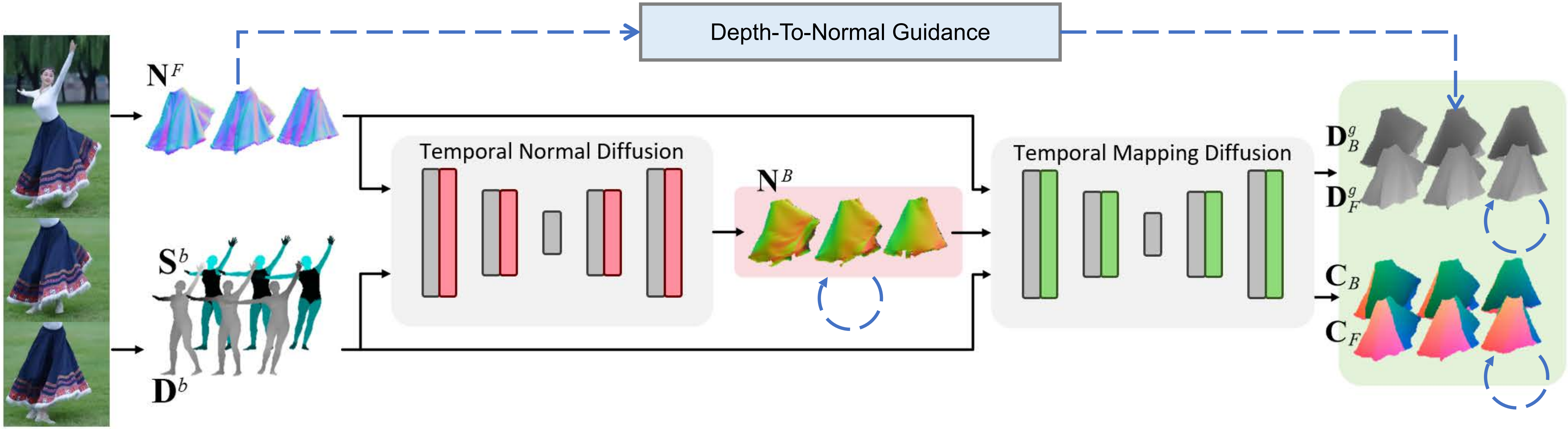
- introduce temporal diffusion models
- enforce geometric and temporal guidance
- temporal consistency guidance, depth-to-normal guidance, interpenetration-aware guidance
- fit the prior to the positional maps with projection-based constraint for reconstruction



From Images to Videos

To handle video cases, we

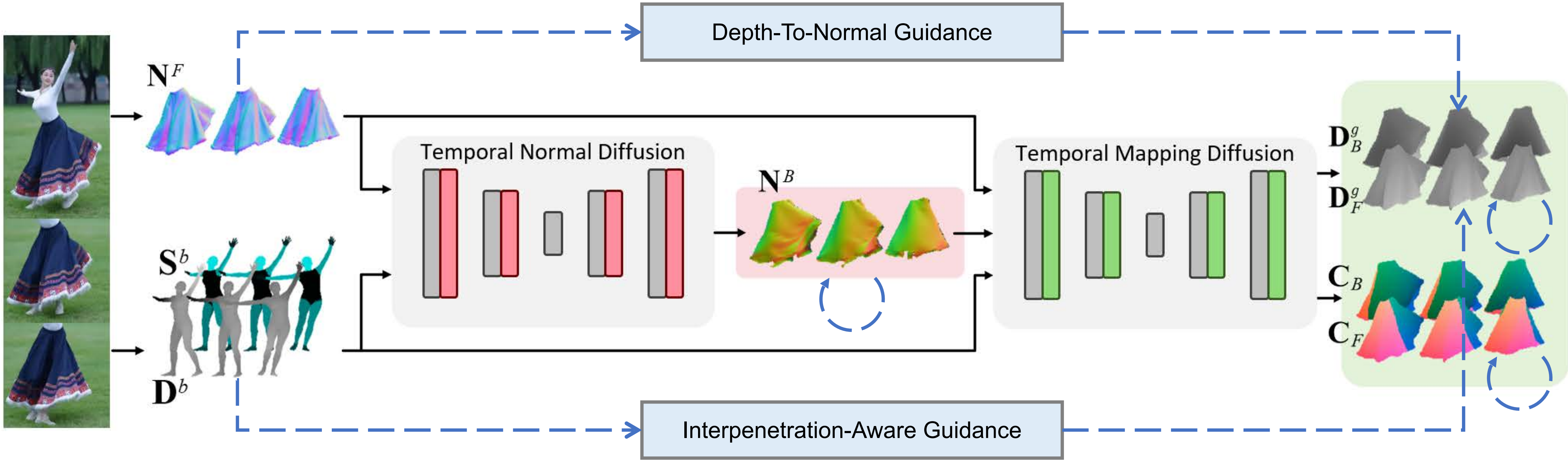
- introduce temporal diffusion models
- enforce geometric and temporal guidance
- temporal consistency guidance, depth-to-normal guidance, interpenetration-aware guidance
- fit the prior to the positional maps with projection-based constraint for reconstruction



From Images to Videos

To handle video cases, we

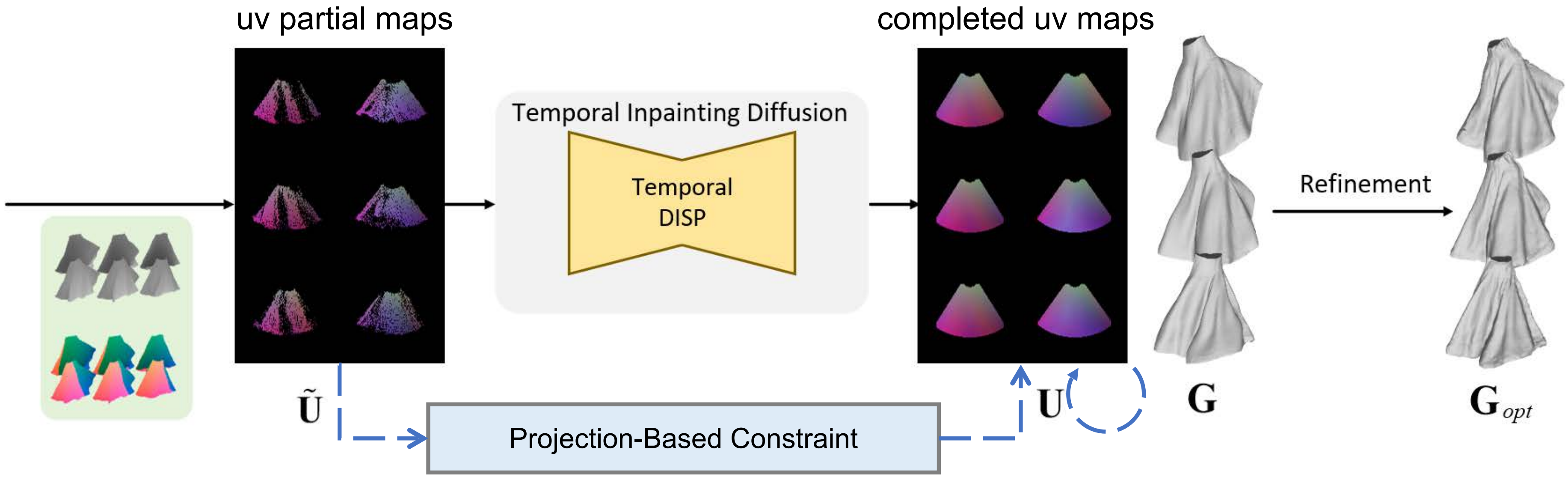
- introduce temporal diffusion models
- enforce geometric and temporal guidance
- temporal consistency guidance, depth-to-normal guidance, interpenetration-aware guidance
- fit the prior to the positional maps with projection-based constraint for reconstruction



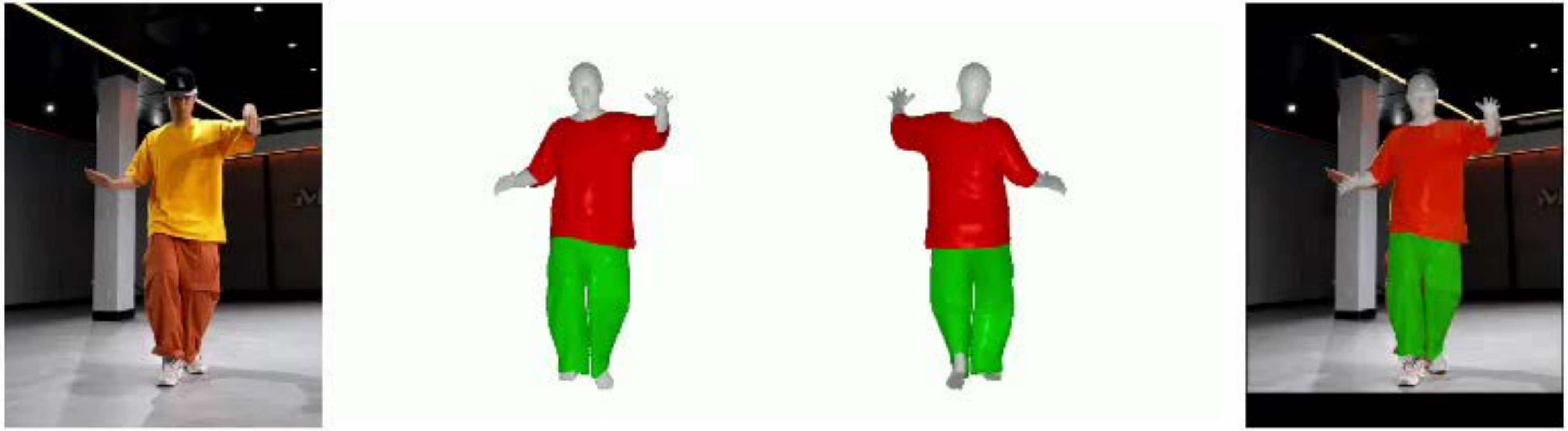
From Images to Videos

To handle video cases, we

- introduce temporal diffusion models
- enforce geometric and temporal guidance
- temporal consistency guidance, depth-to-normal guidance, interpenetration-aware guidance
- fit the prior to the positional maps with projection-based constraint for reconstruction



Recovered Garments from Videos



Input Front-view Back-view Projection

What about Long Robes?

What about Long Robes?

- Quite good but not quite right when seen from the side.
- From a physical point of view, the 3D pose is not realistic.

➡There still is work to do!

PhD anyone?