Invertible Neural Skinning

Yash Kant, Aliaksandr Siarohin, Riza Alp Guler, Menglei Chai, Sergey Tulyakov, Igor Gilitschenski

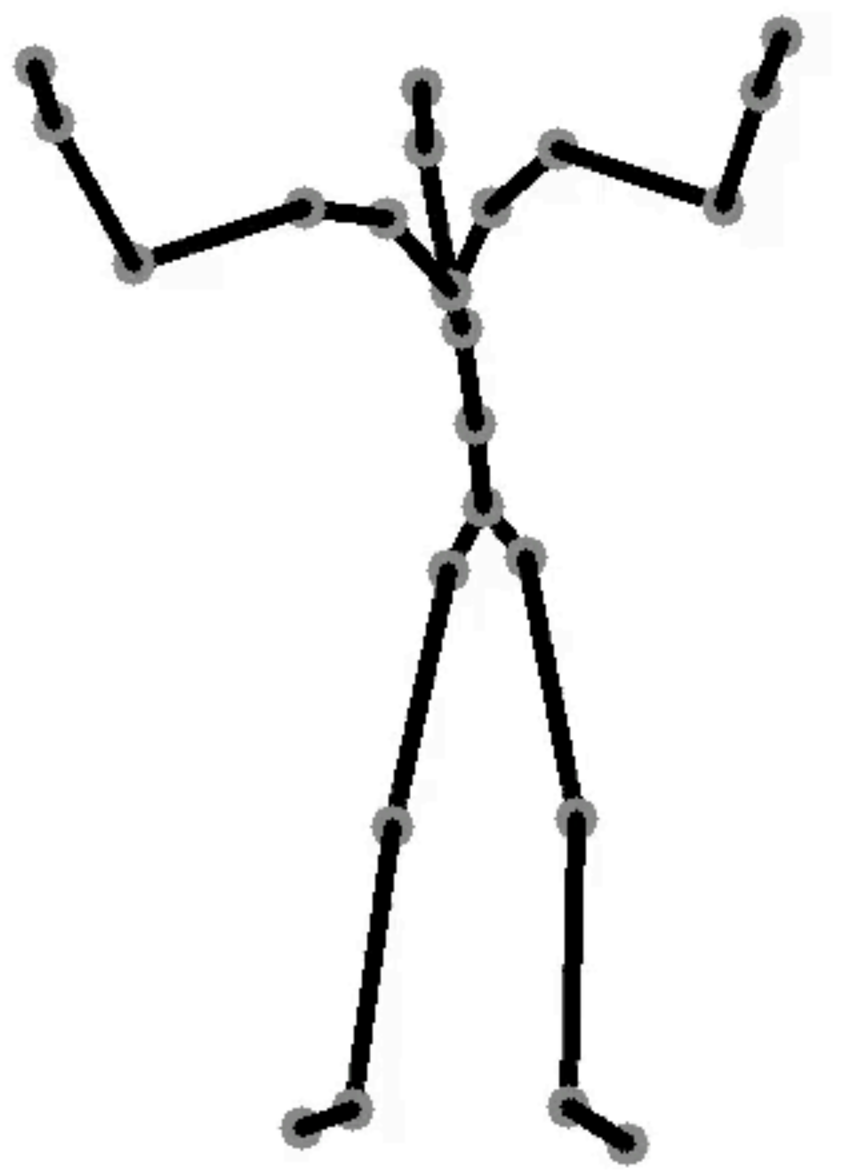
Conference on Computer Vision and Pattern Recognition (CVPR), 2023

University of Toronto
Snap Research

Reposing Task.

Given: Sequence of **3D scans** (meshes) and **SMPL-fitted poses.**

3D scans

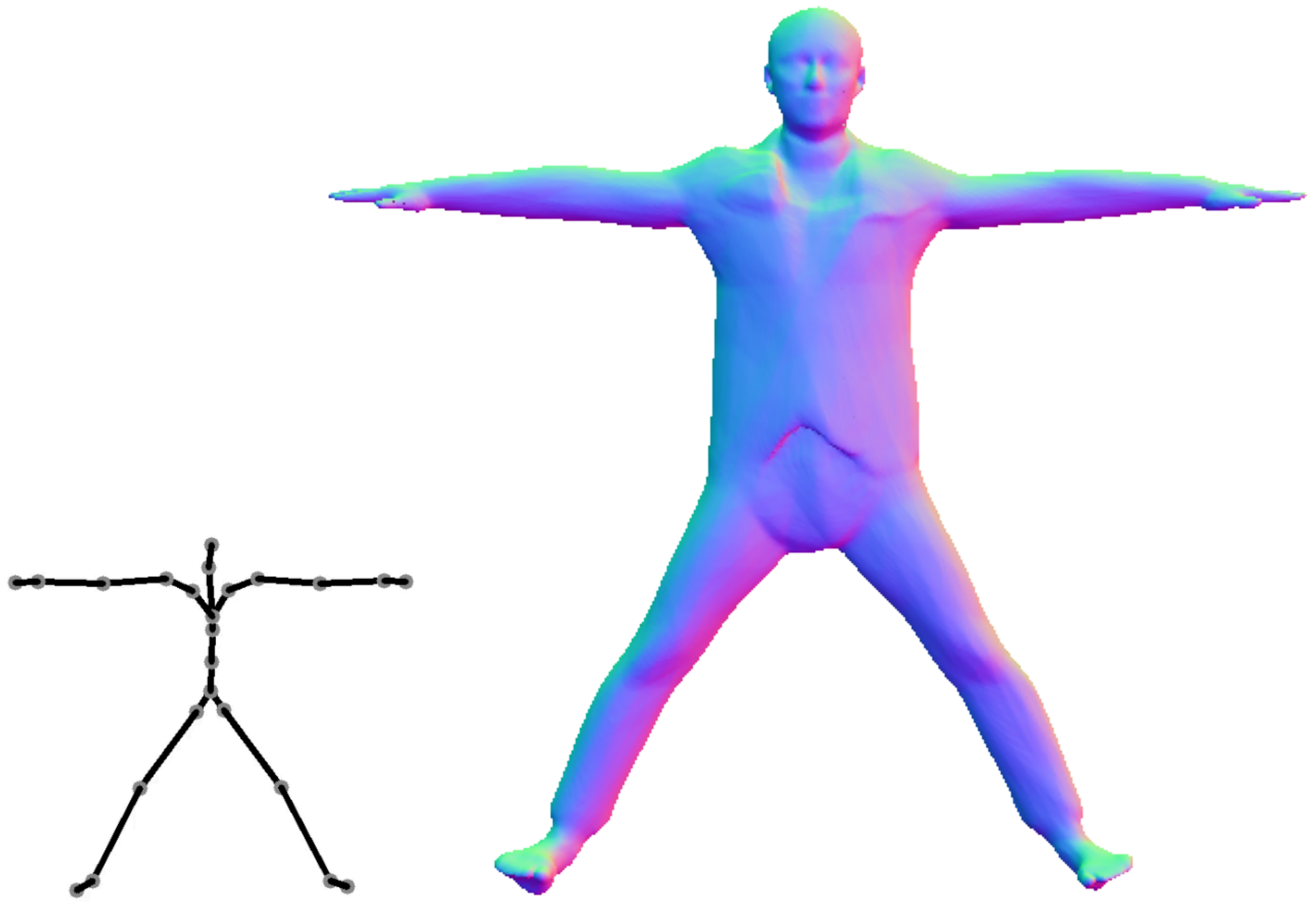


SMPL-fitted poses

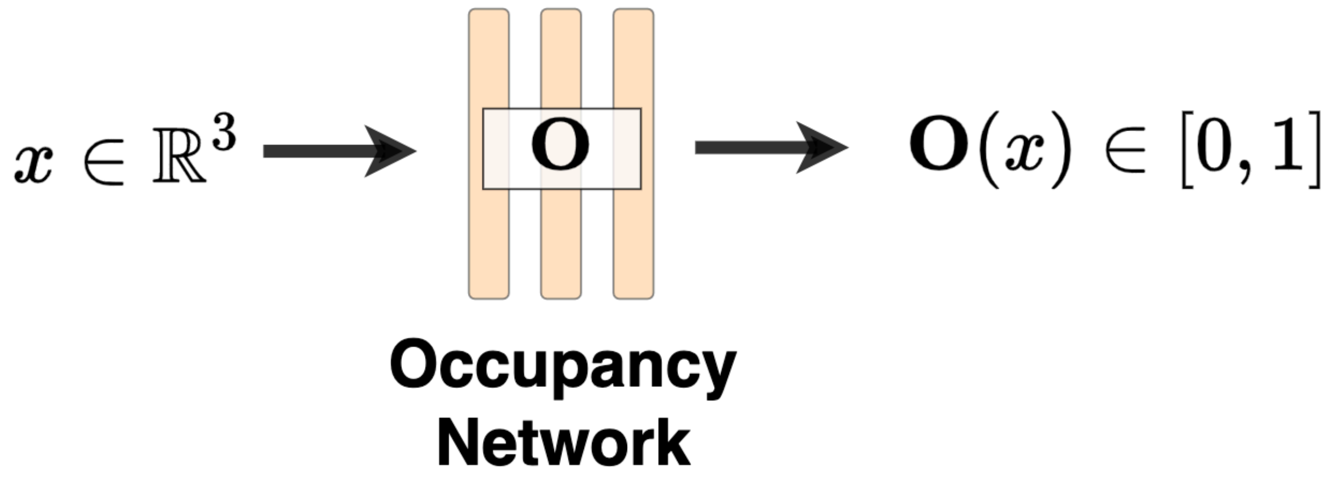
Canonical Space. [Background]

We first define a canonical space where the subject pose is fixed.

And we learn **canonical shape** implicitly using an occupancy network.



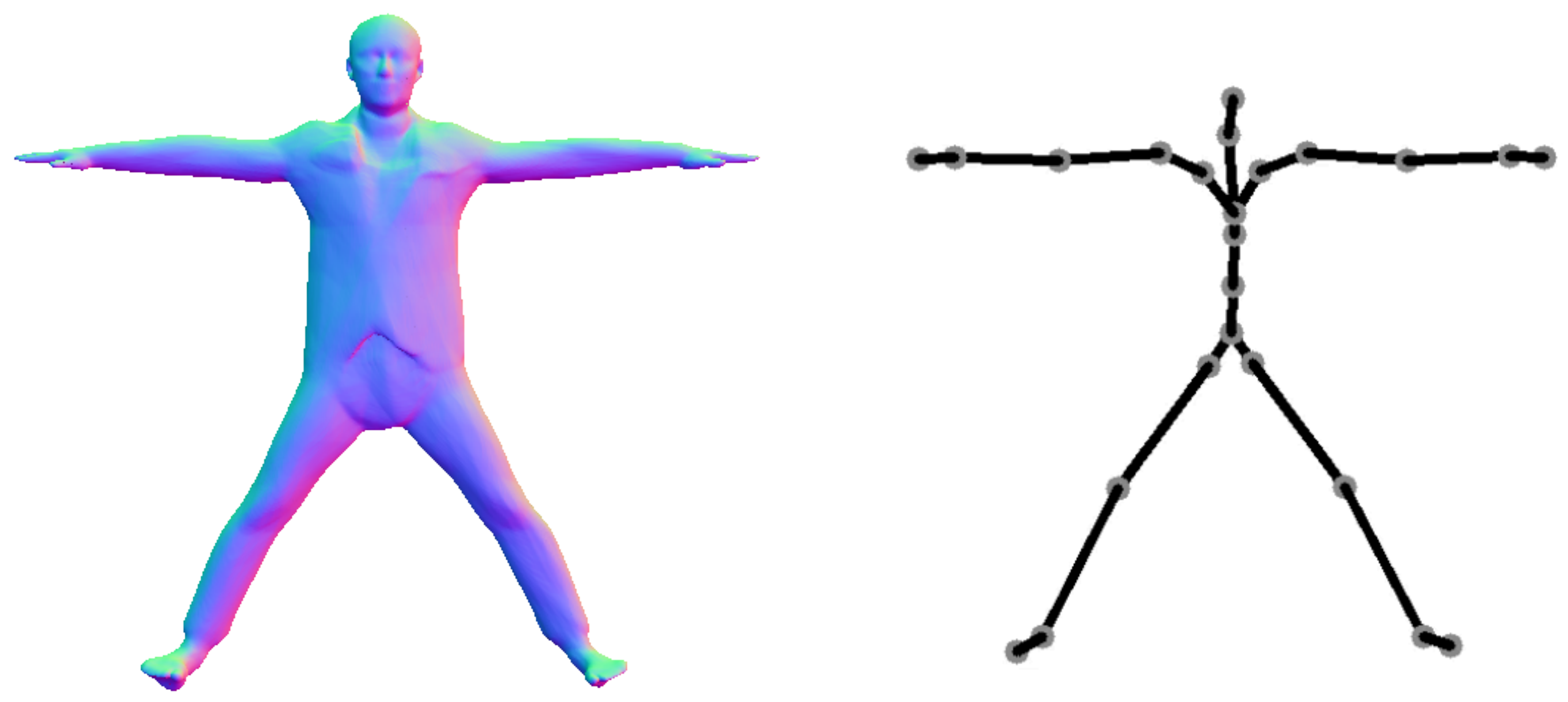
Canonical Representation



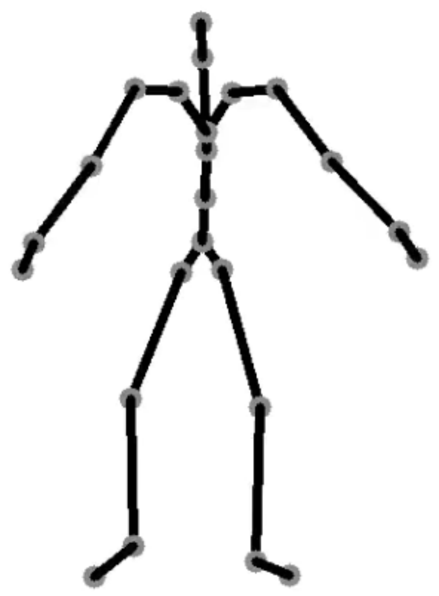
Skinning. [Background]

Linear Blend Skinning (LBS) is a way to animate 3D surfaces based on relative bone configurations (canonical to deformed).

Given



Canonical Representation



Novel Target Pose

$$\mathbf{v}'_i = \sum_{j=1}^{m} w_{i,j} \mathbf{T}_j \mathbf{v}_i$$

Output (Deformed Space)

LBS Animated

Skinning. [Background]

For each point v_i on the surface, we define a set of weights w_{ij} that defines how much the j^{th} bone contributes to its movement.

$$\mathbf{v}'_i = \sum_{j=1}^{m} \underbrace{w_{i,j}\mathbf{T}_j}_{j^{th}\text{ bone}} \overbrace{\mathbf{v}_i}^{i^{th}\text{ point}}$$

Skinning. [Background]

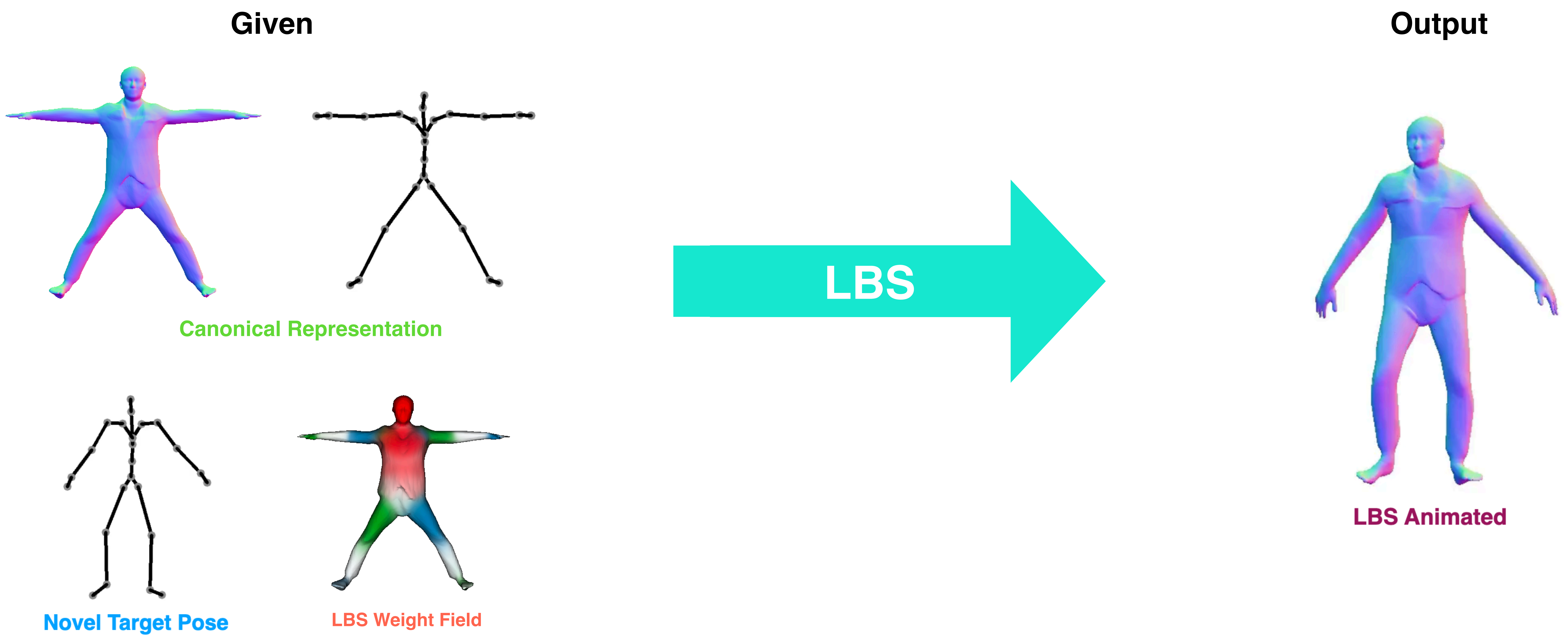
And all weights for an individual point sum to one.

$$\mathbf{v}'_i = \sum_{j=1}^{m} \underbrace{\boxed{w_{i,j} \mathbf{T}_j}}_{j^{th}\ \textbf{bone}} \overbrace{\boxed{\mathbf{v}_i}}^{i^{th}\ \textbf{point}}$$

$$\sum_{j=1}^{|B|} w_{ij} = 1, \forall i$$

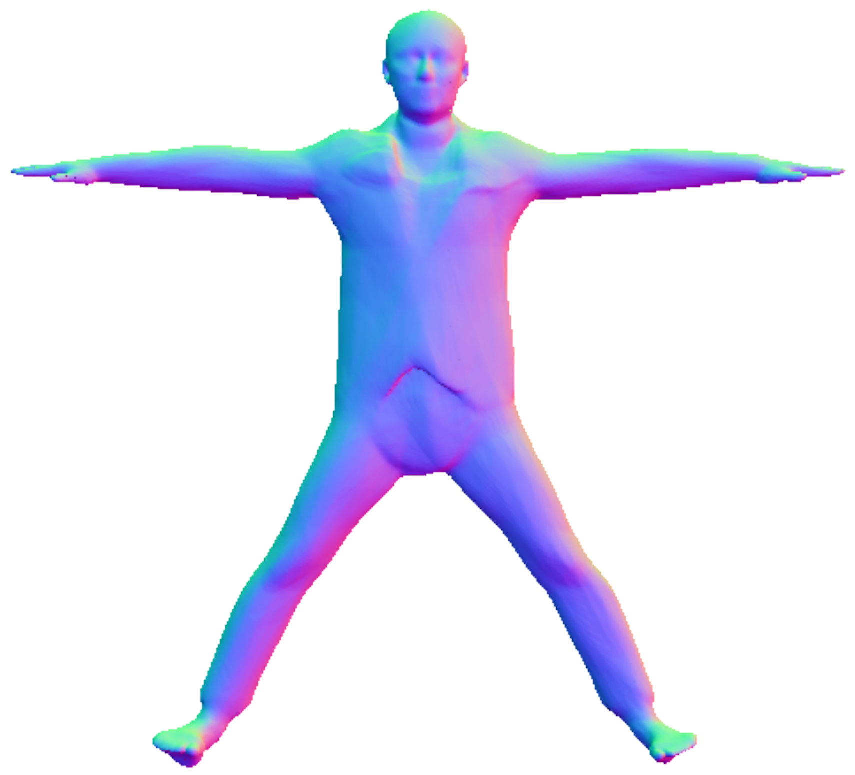
Skinning. [Background]

To use LBS, we learn an extra MLP that predicts LBS weights for every point.

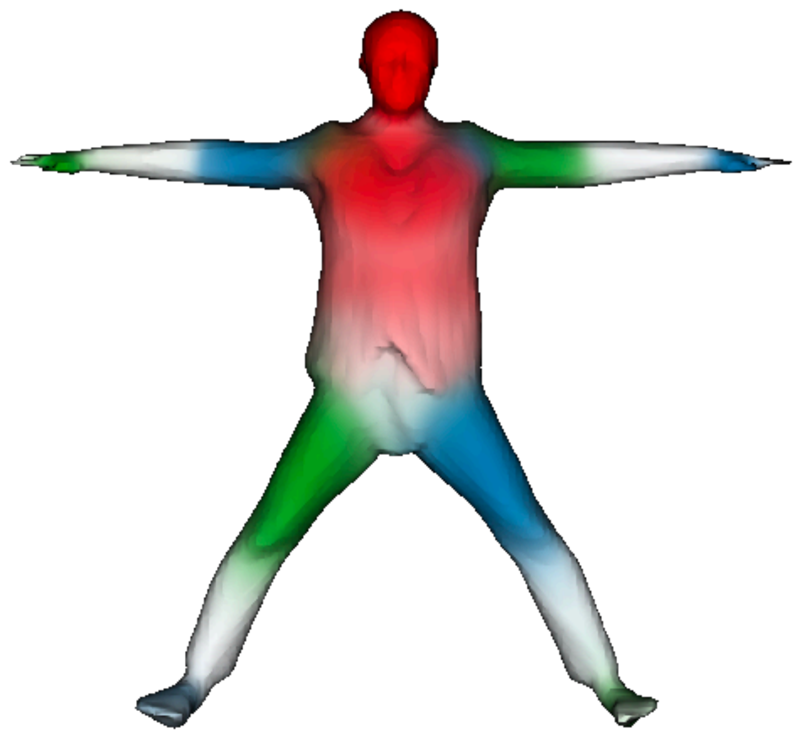


LBS Shortcomings. [Background]

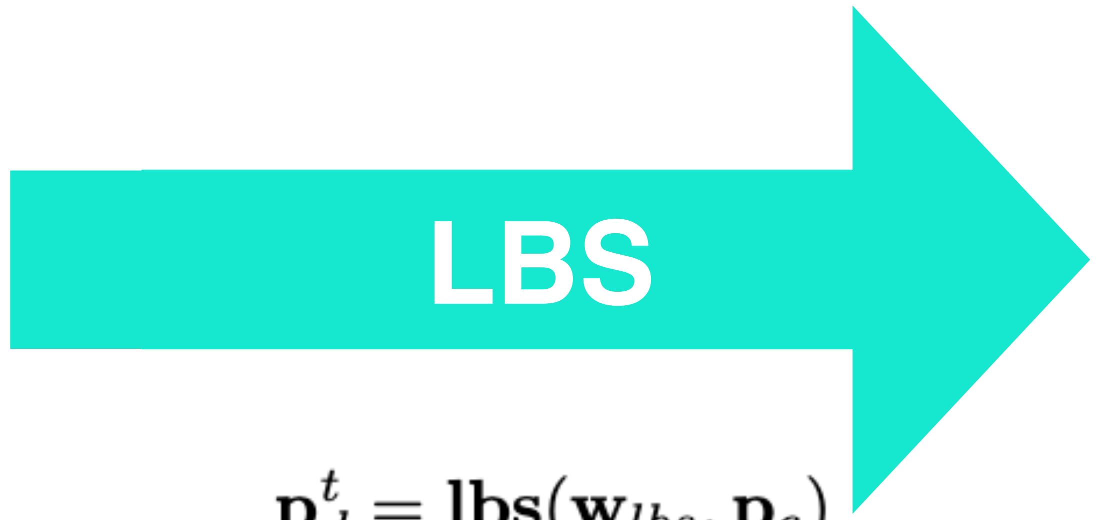
But LBS **cannot capture** **non-linear deformations** of clothes and body tissue well.



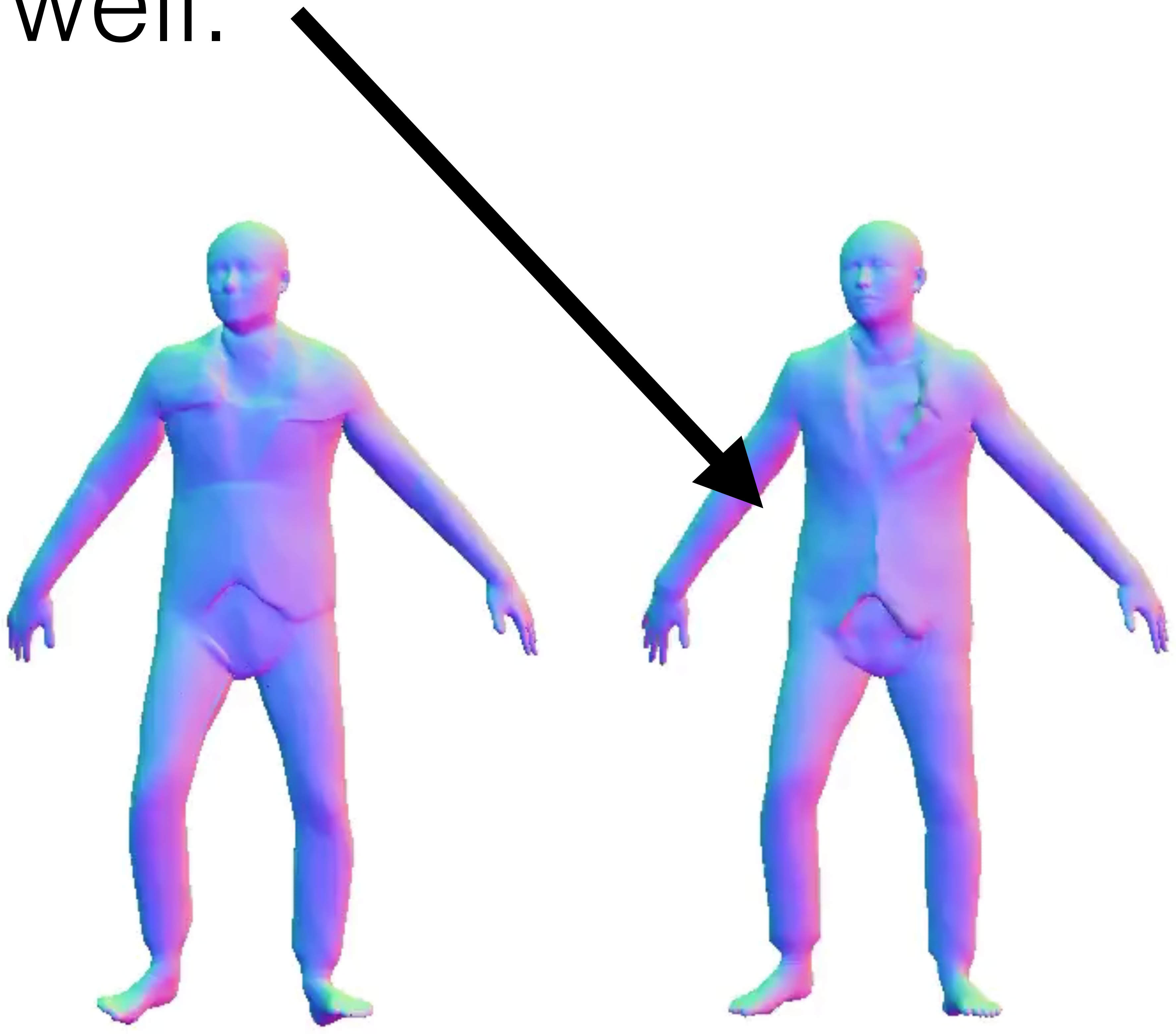
Canonical Representation



LBS Weight Field

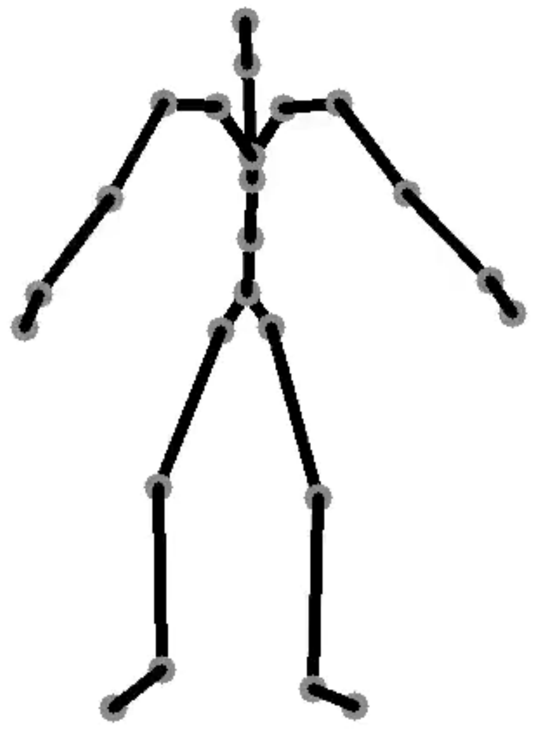


$$\mathbf{p}_d^t = \mathbf{lbs}(\mathbf{w}_{lbs}, \mathbf{p}_c)$$



LBS Animated

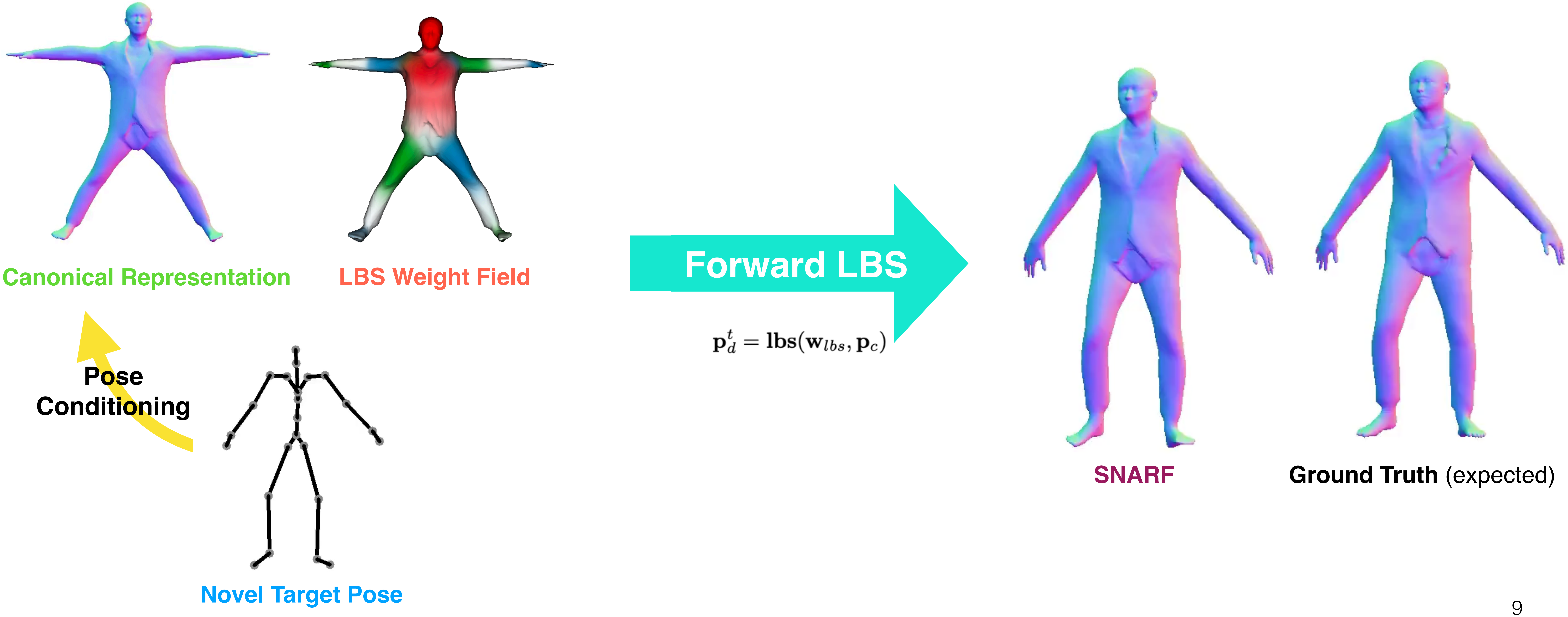
Ground Truth (expected)



Novel Target Pose

Fixing LBS. [Background]

To overcome this, prior works (eg. SNARF) **condition** canonical representation **on target pose**.



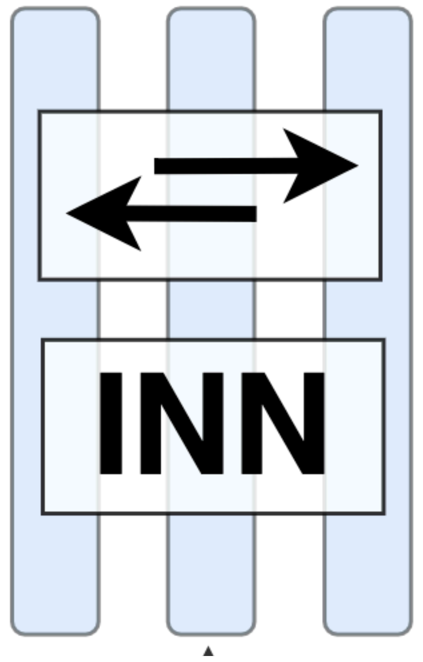
Drawbacks of pose-conditioned canonical space.

- **For each new pose,** we have to **extract a new mesh** first and then animate it. [**expensive operation**]
- The **same subject has different meshes** (vertices, faces) for different poses. [**no correspondences**]

SNARF

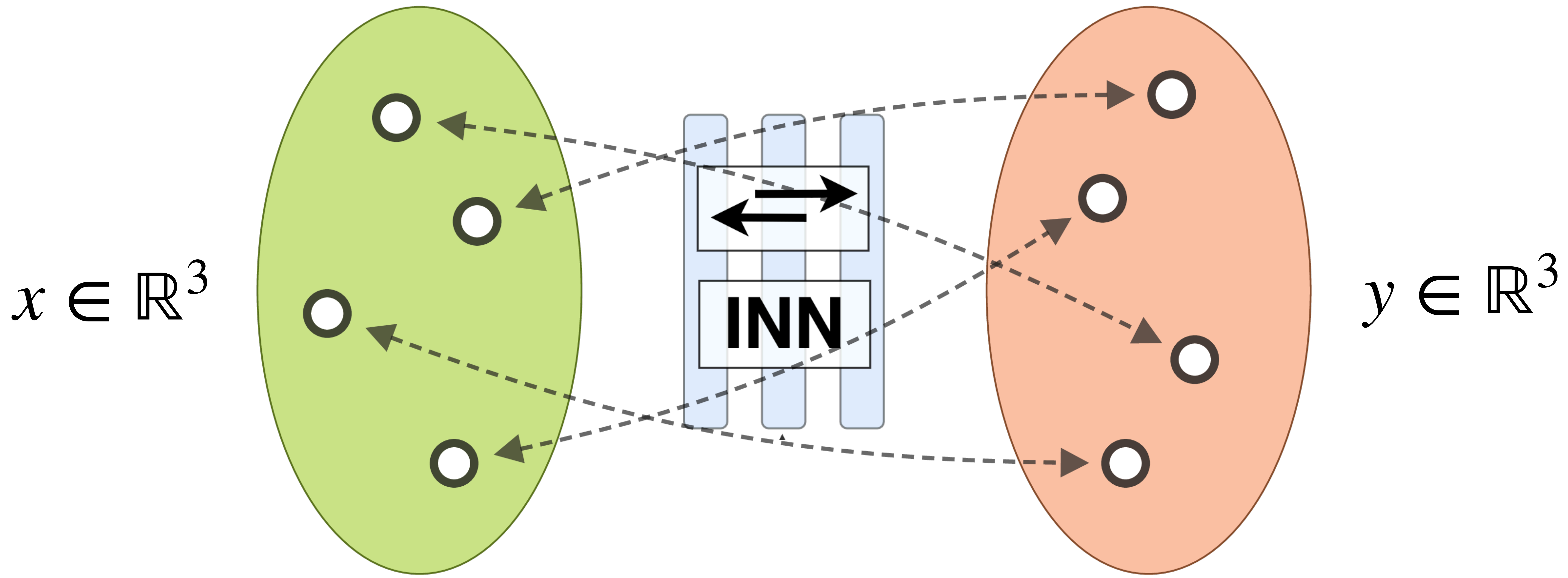
Invertible Neural Skinning [Approach]

Our core contribution is the use of **Invertible Neural Network (INN)** in the reposing pipeline.



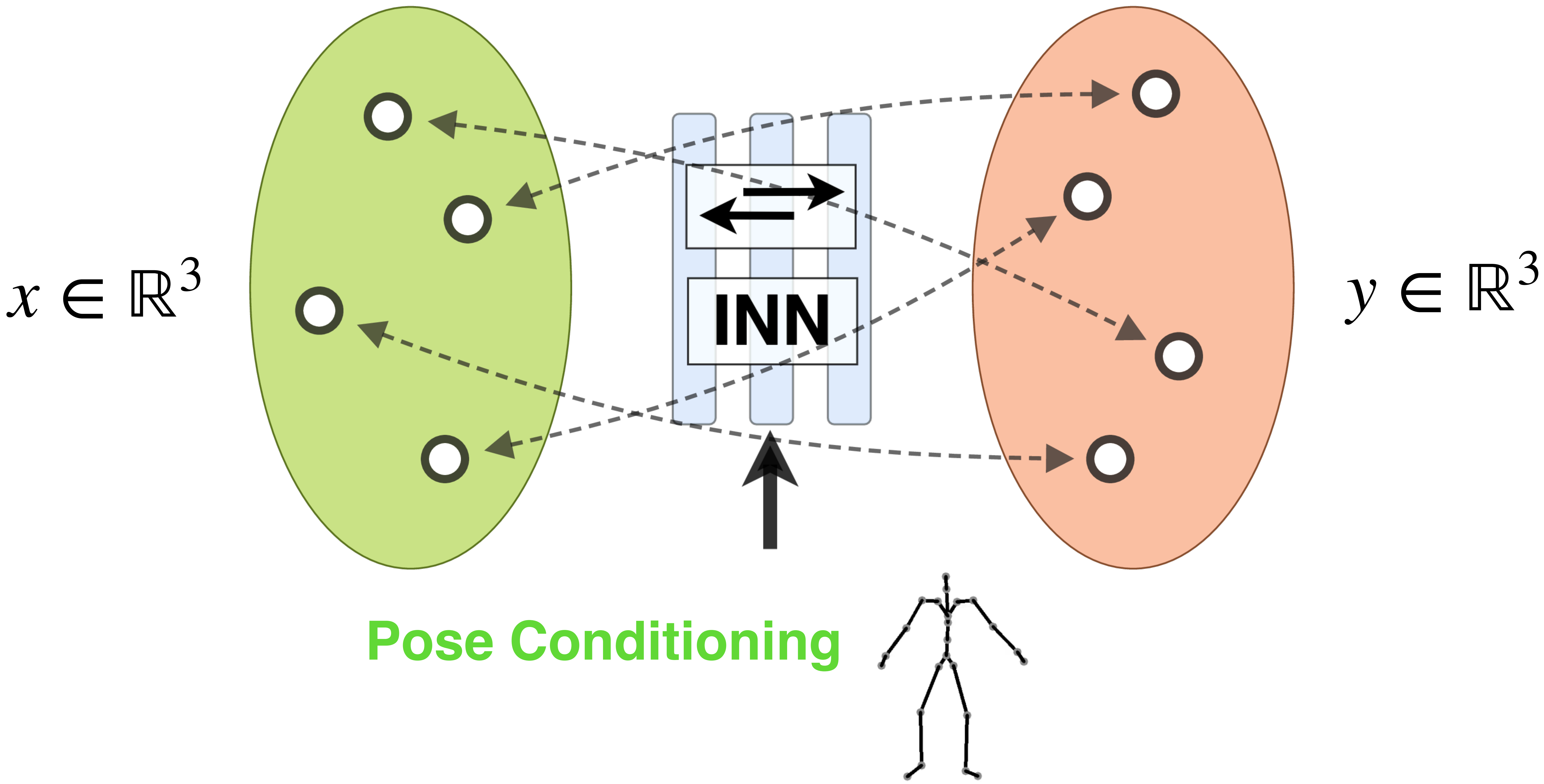
Invertible Neural Skinning [Approach]

An **Invertible Neural Network (INN)** defines a bijective mapping between its input and output spaces.



Invertible Neural Skinning [Approach]

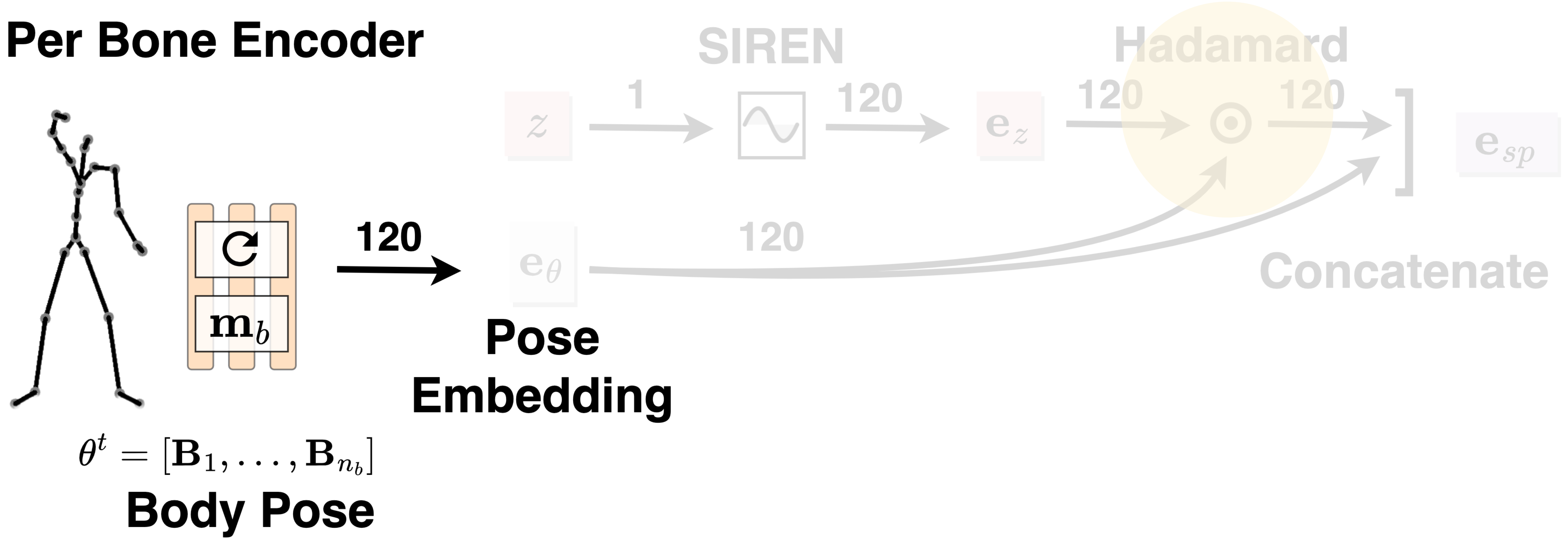
We introduce **Pose-conditioned INN (PIN)** to handle non-linear deformations, such as those of clothes.



Bone Pose Encoder. [Approach]

Our second contribution is the bone pose encoder.

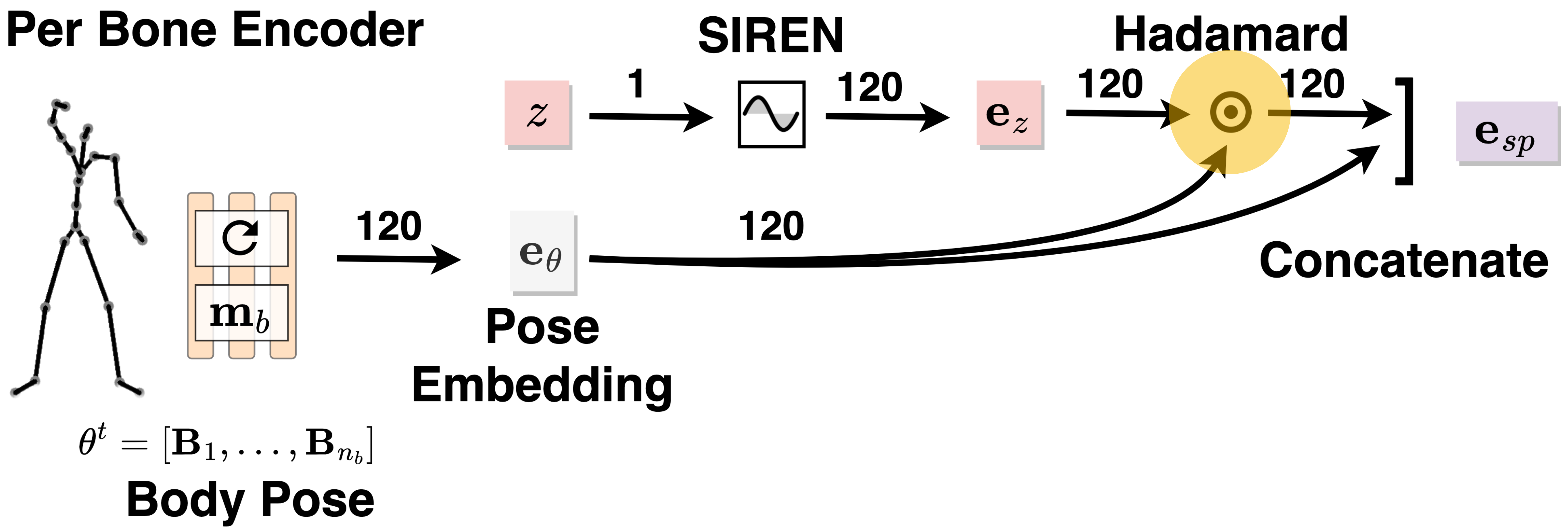
We encode each bone pose (rotation and translation) separately, and concatenate them.



Bone Pose Encoder. [Approach]

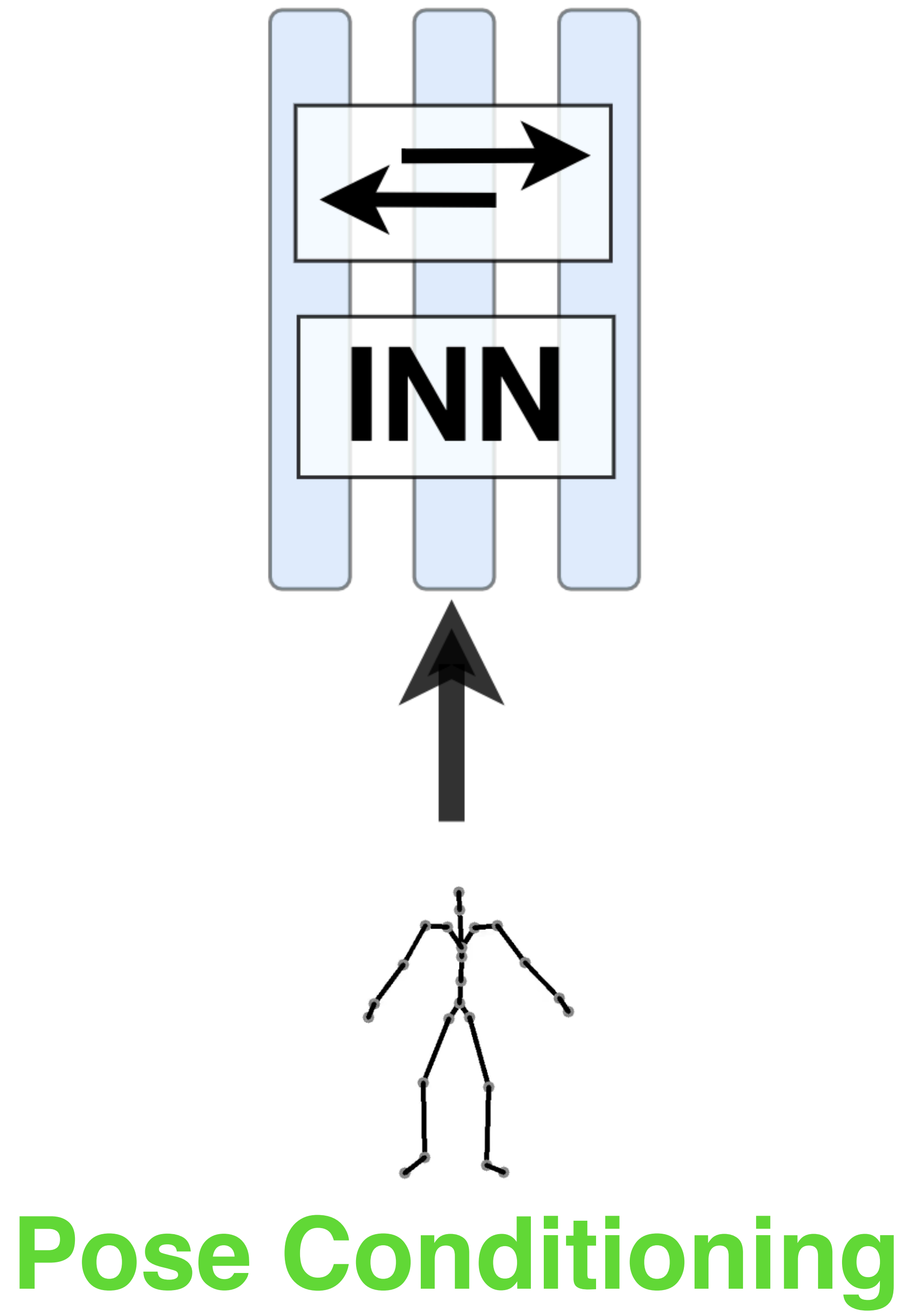
We take a dot product of this pose embedding with the spatial embedding.

This ensures that deformations in INN only occur when the pose-embedding is non-zero.



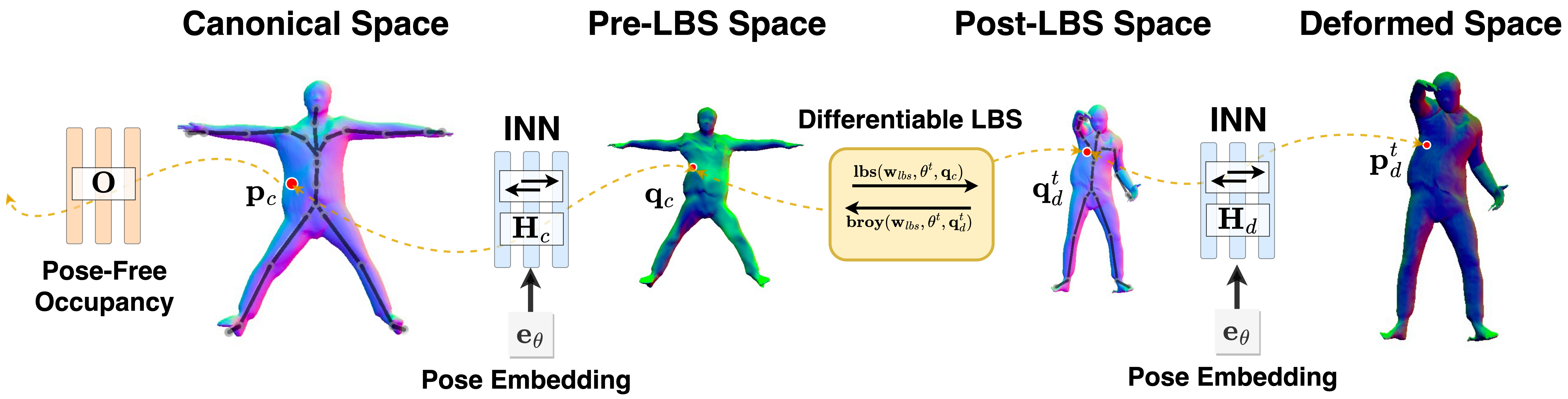
Pose-conditioned INN. [Approach]

Using bone pose encoder we build Pose-conditioned INN, and use it to model non-linear deformations.



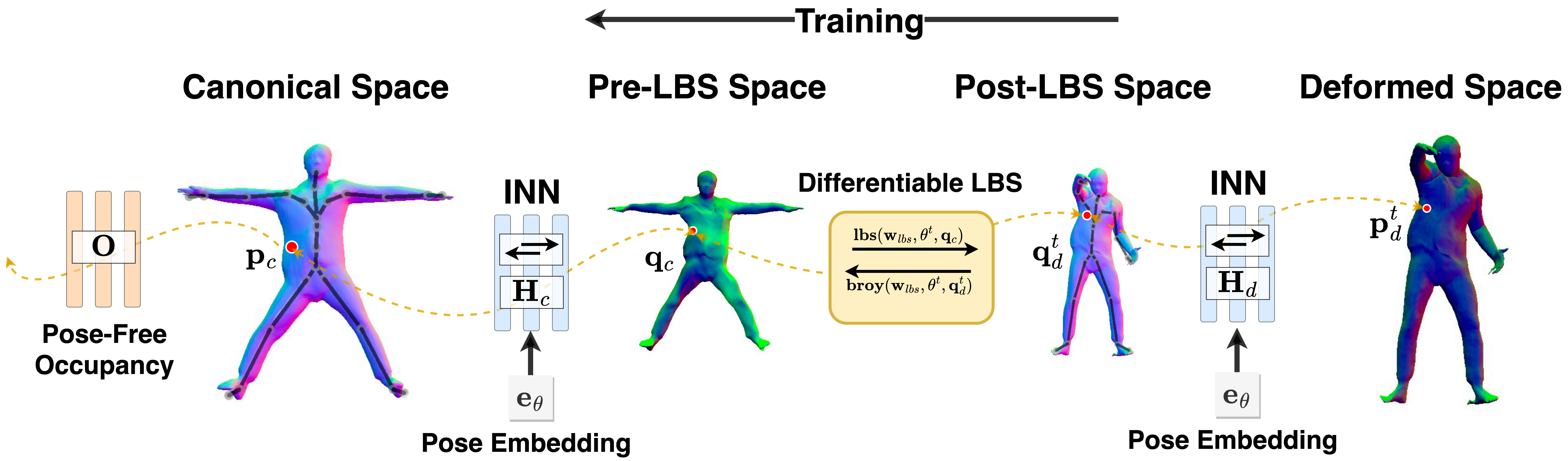
Invertible Neural Skinning [Overall]

We chain **two PINs** around an **LBS** block to build final reposing pipeline.



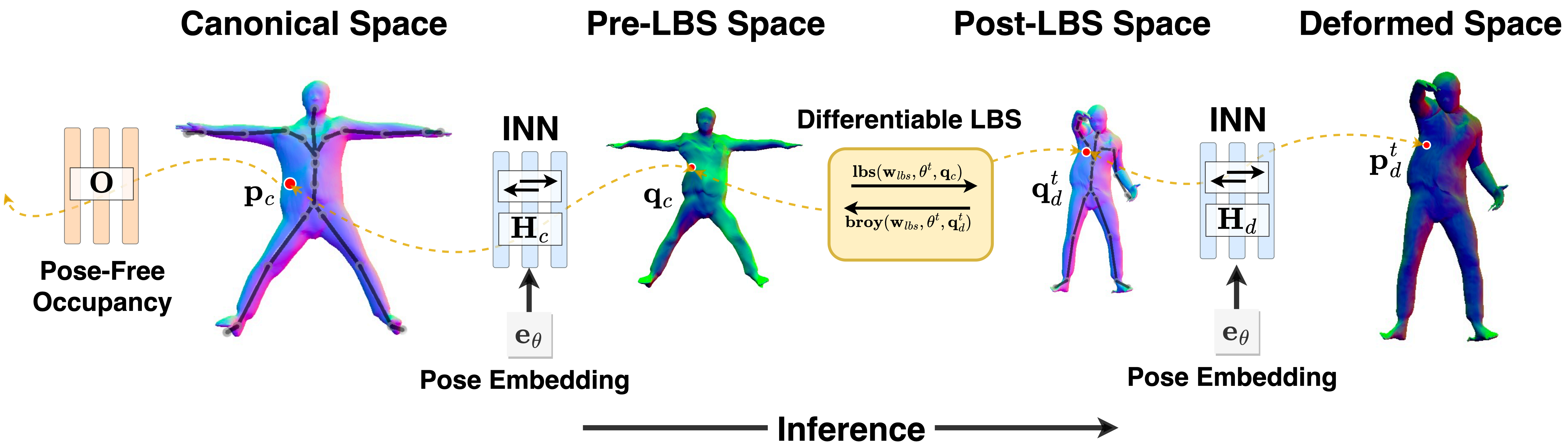
Invertible Neural Skinning [Overall]

Training: Sample points in the deformed space, and train our network to predict its occupancy [0/1], use BCE loss.



Invertible Neural Skinning [Overall]

Inference: First extract a mesh in canonical space only once, and repose it using learned LBS and PINs.



Metrics.

Bounding Box IoU: How many **points sampled uniformly** in space have correct occupancy?

Surface IoU: How many **points sampled around the ground truth surface** have correct occupancy?

Results.

We match/outperform previous methods, and baselines on both metrics on CAPE (clothed human dataset).

		IoU Surface					IoU Bounding Box				
Subject	Clothing	AVG-LBS	FIRST-LBS	SNARF	SNARF-NC	INS (ours)	AVG-LBS	FIRST-LBS	SNARF	SNARF-NC	INS (ours)
Average		65.01%	57.41%	72.24%	66.89%	**73.13%**	65.12%	57.5%	72.17%	66.78%	**73.19%**

Results.

INS does not require mesh extraction at each step, so it is an **order of magnitude faster** than baseline.

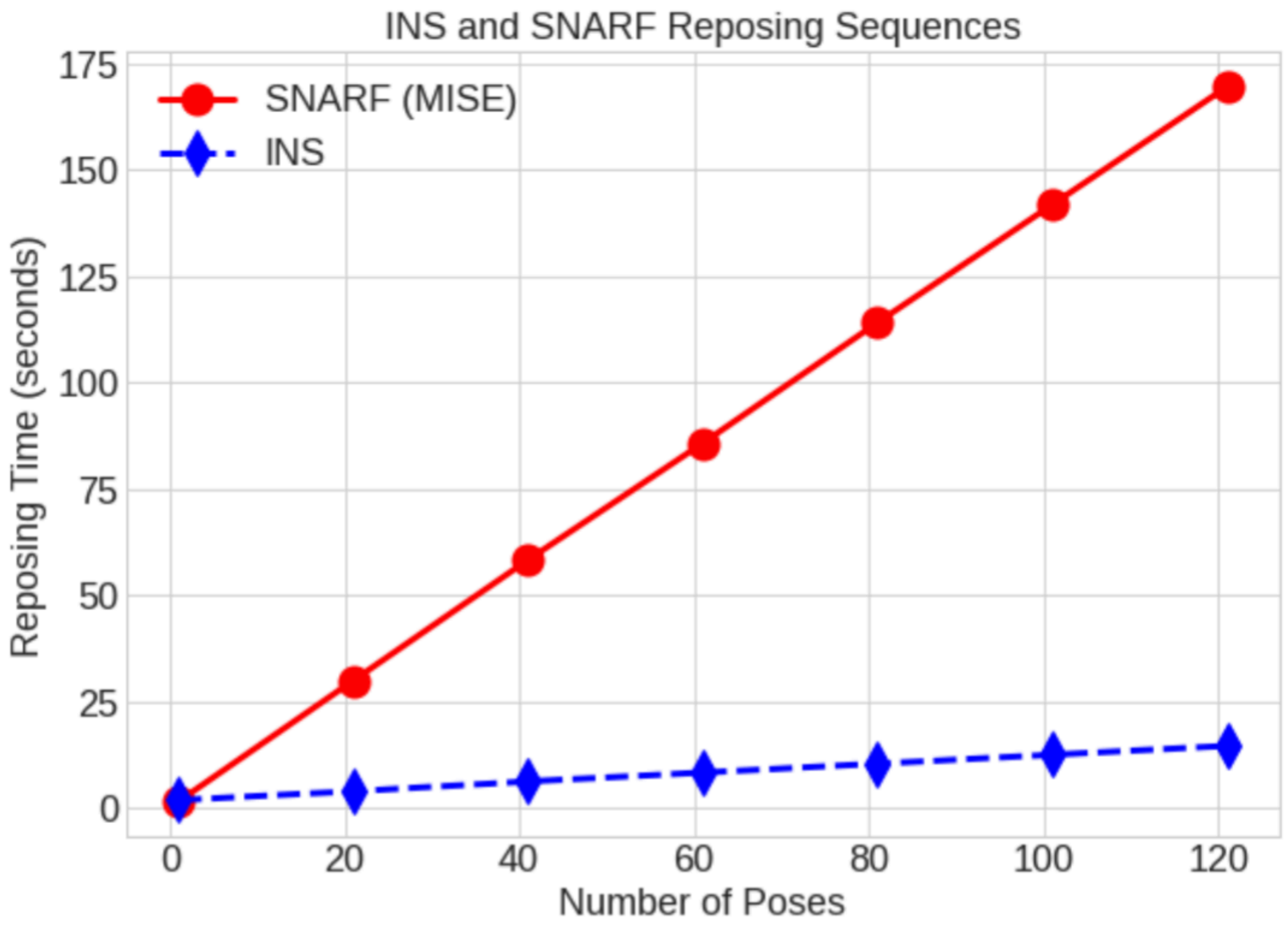


Figure 5. **Reposing time comparison between INS and SNARF** We show the time taken by SNARF vs INS for reposing a mesh extracted at 128^3 resolution across 125 different target poses. INS performs reposing an order of magnitude faster than SNARF.

Ablation Study.

Not using bone-pose multiplication, leads to a huge drop in performance.

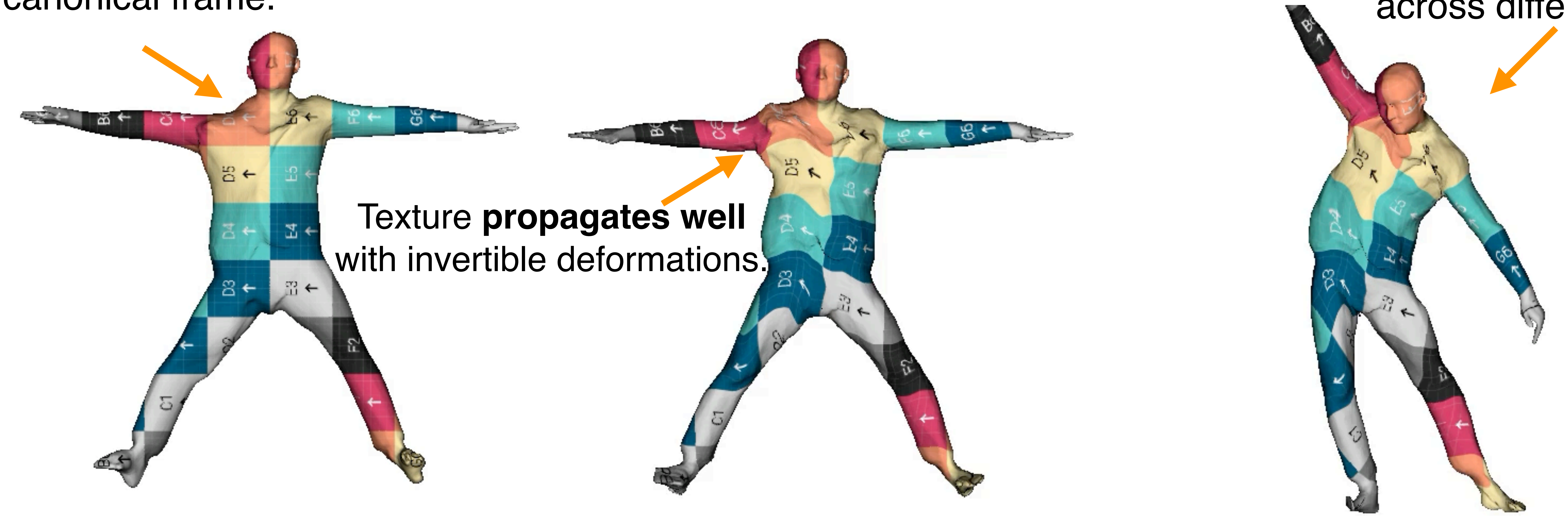
#	Ablation	IoU Surface (%)	IoU Bounding Box(%)
1	INS(vanilla)	**72.83**	**72.69**
2	w/o Pose Mul.	$61.94_{-\mathbf{10.89}}$	$62.00_{-\mathbf{10.69}}$
3	w/o SIREN	$69.67_{-3.16}$	$69.57_{-3.12}$
4	w/o Rotation	$71.91_{-0.92}$	$71.87_{-0.82}$
5	w/o $\mathbf{H}_d$	$72.66_{-0.17}$	$72.58_{-0.11}$
6	w/o $\mathbf{H}_c$	$67.89_{-\underline{4.94}}$	$67.81_{-\underline{4.88}}$
7	w/o **LBS**	$40.79_{-32.04}$	$40.65_{-32.04}$

Table 3. **Ablation Table.** We perform an ablation study of INS

Qualitative Results [INS].

Texture applied ***only*** to the fixed canonical frame.

Texture **does not overflow** across different regions.



Texture **propagates well** with invertible deformations.

INS (canonical space)

INS deformations by $\mathbf{H}_c$

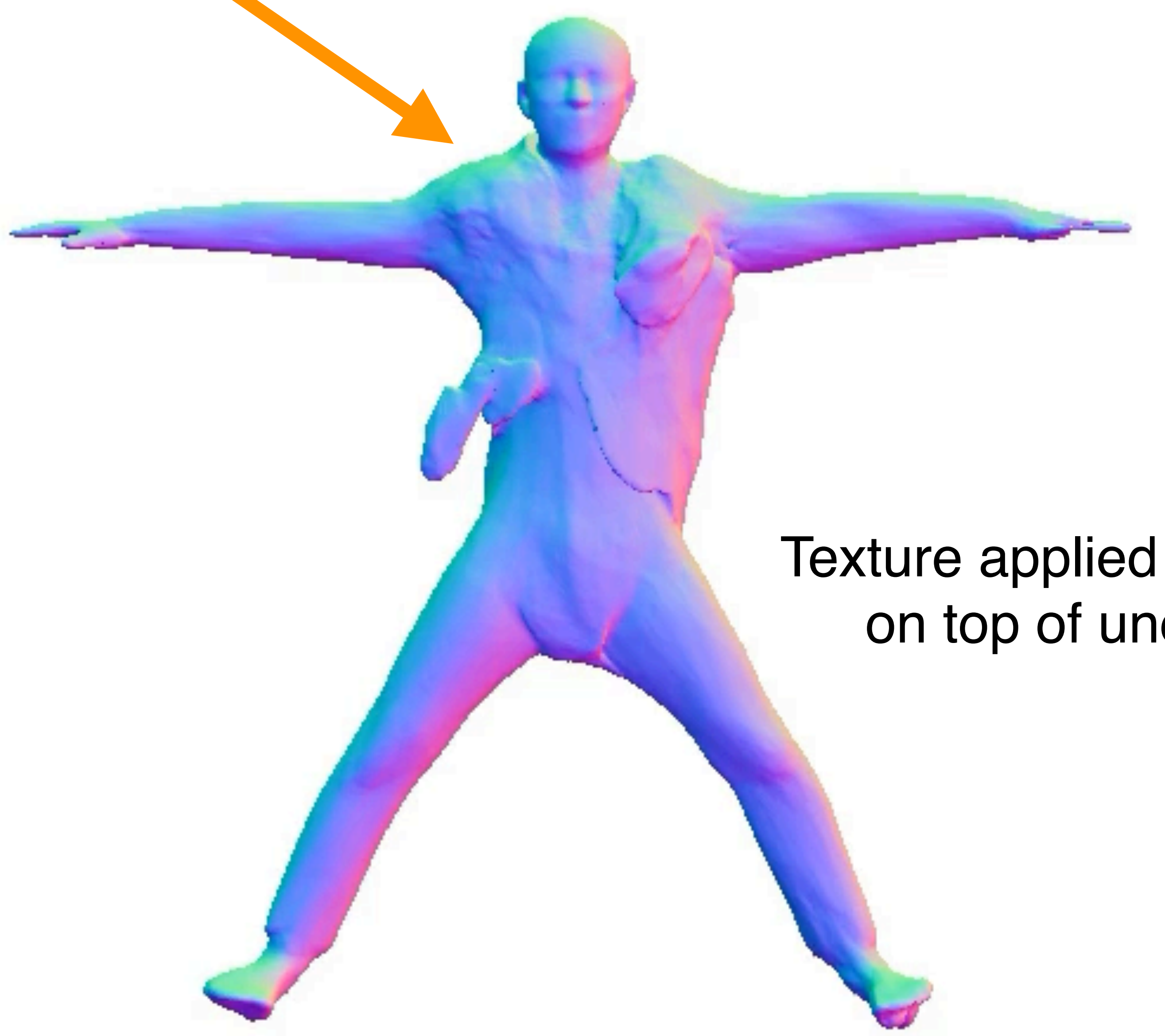
INS (final output)

Texture Propagation using INS — Fast and Consistent

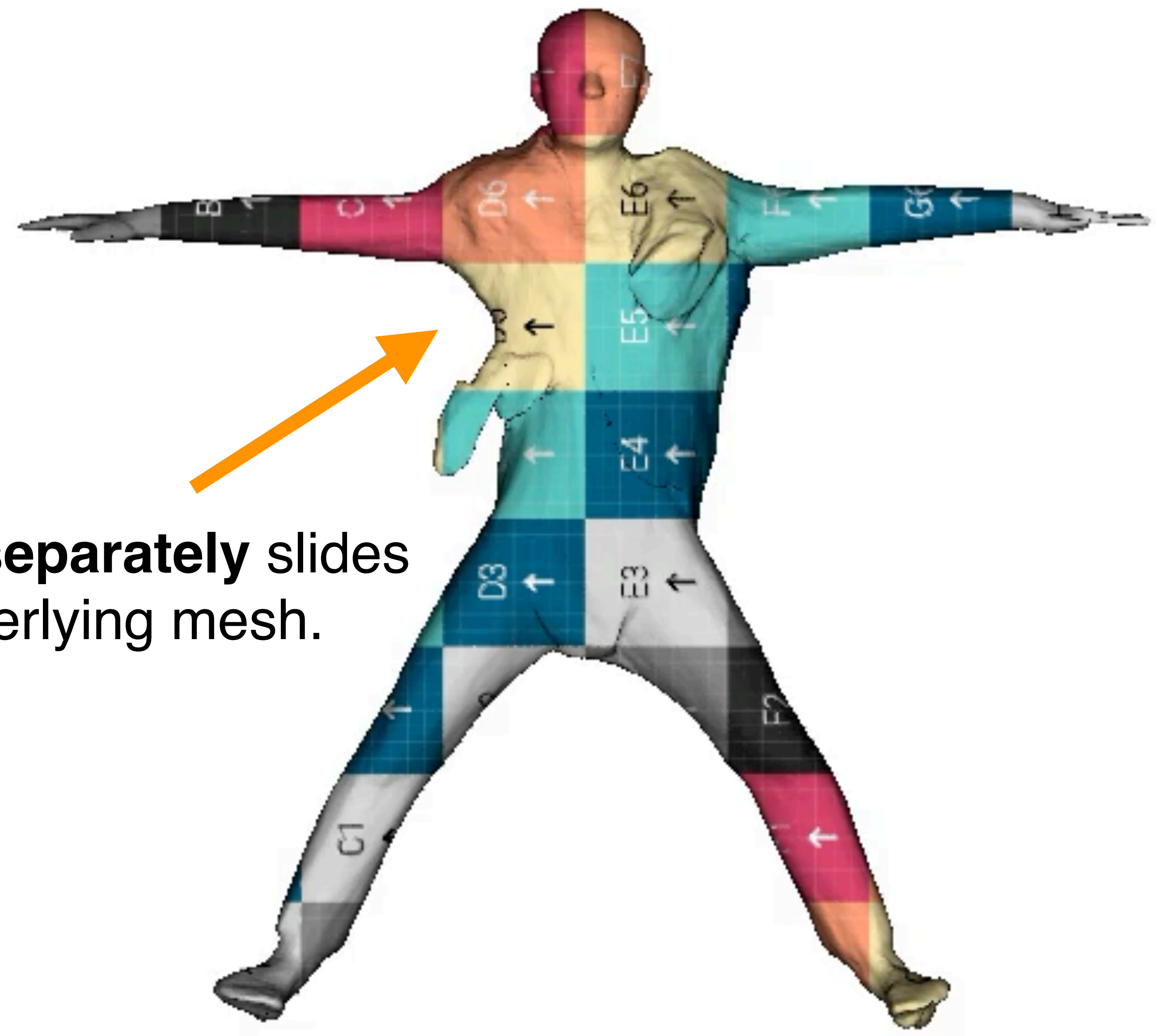
Qualitative Results [SNARF].

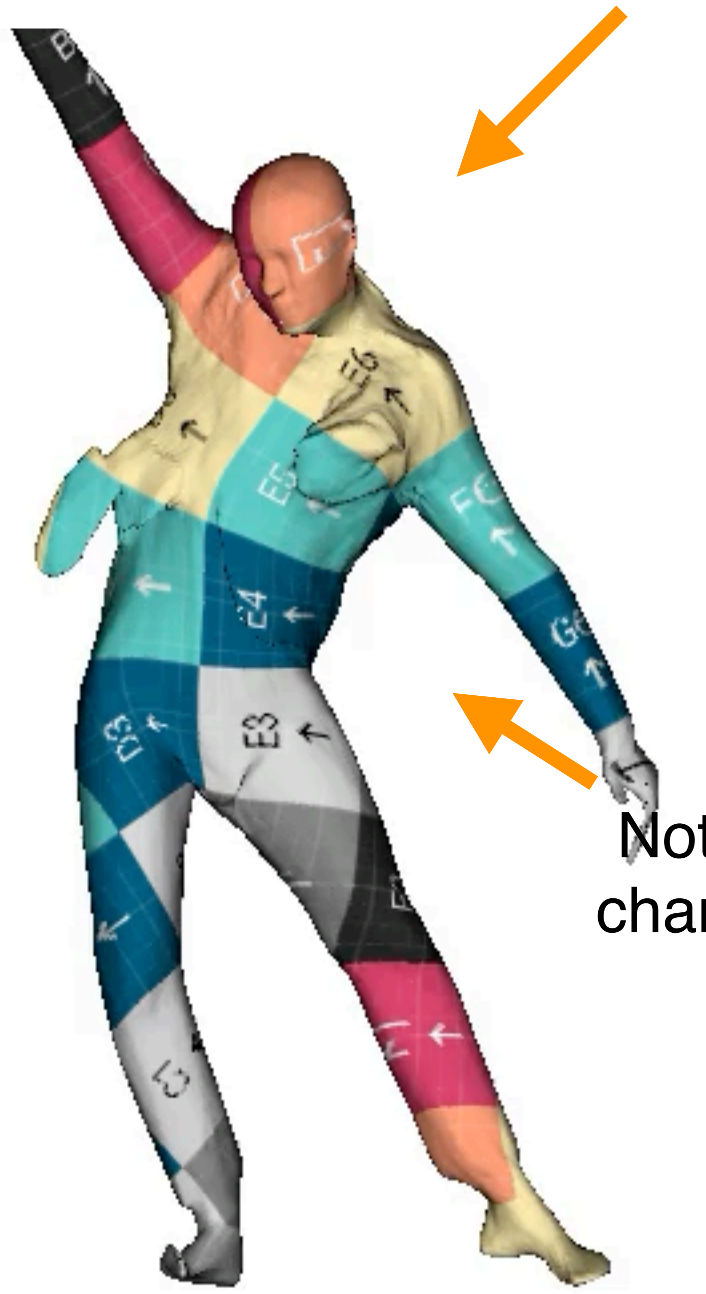
Every frame (mesh) in canonical space has different topology

Causes **jittery artifacts** as frames have inconsistent textures.



Texture applied **separately** slides on top of underlying mesh.





Notice the blazer outline changing colors (E3—E4)

SNARF (canonical space)

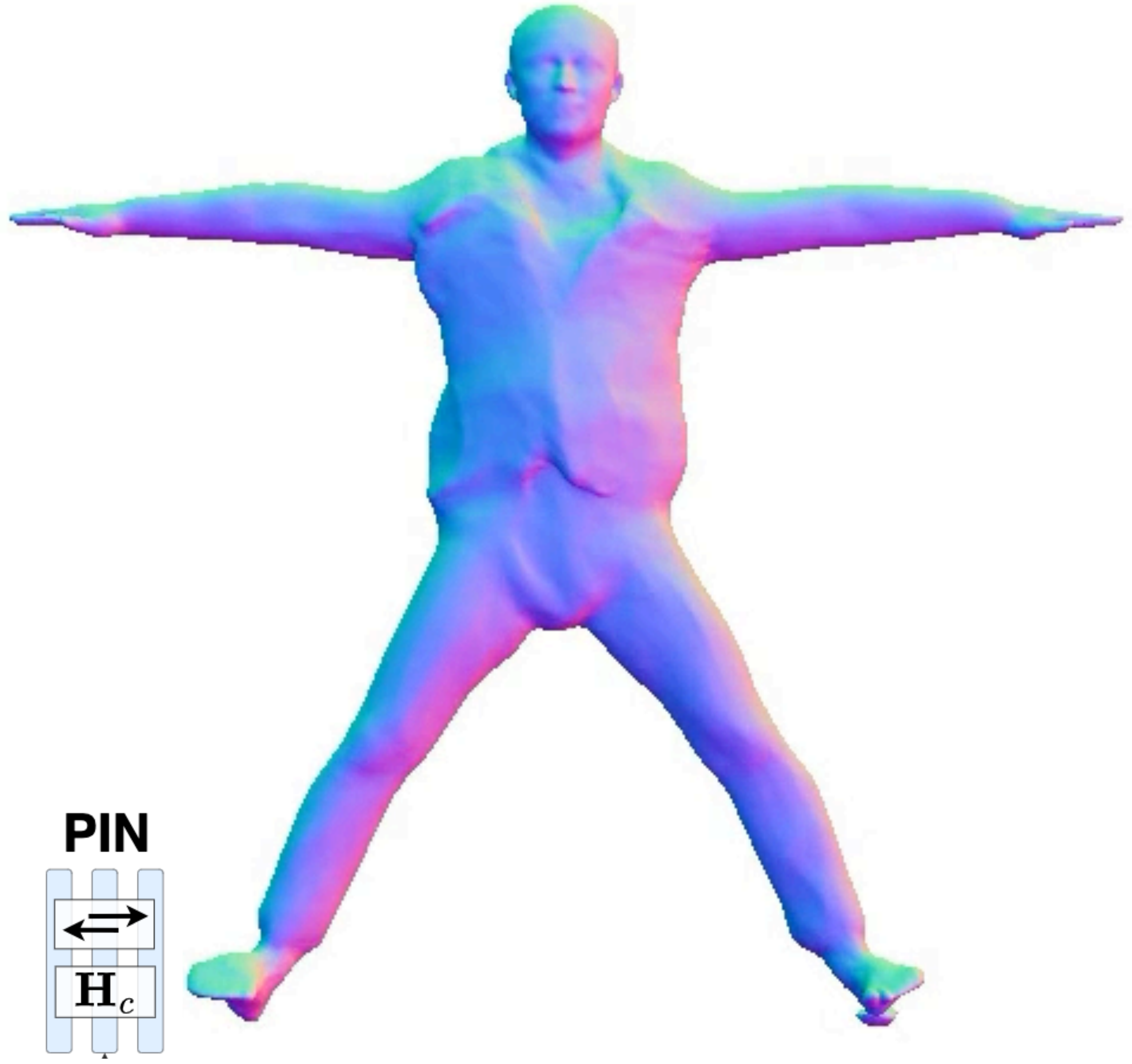
Per-frame Texturing (canonical)

SNARF (jittery output)

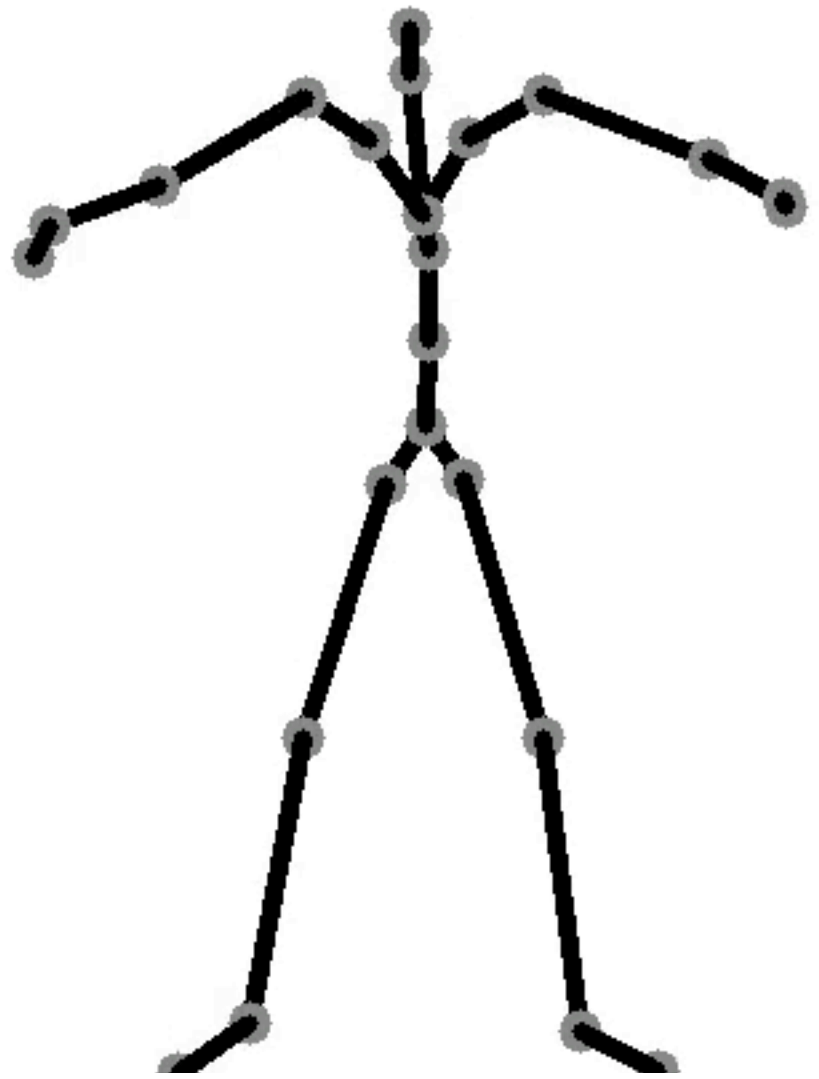
Texture Propagation in SNARF — Slow and Jittery

Pose-varying INS deformations.

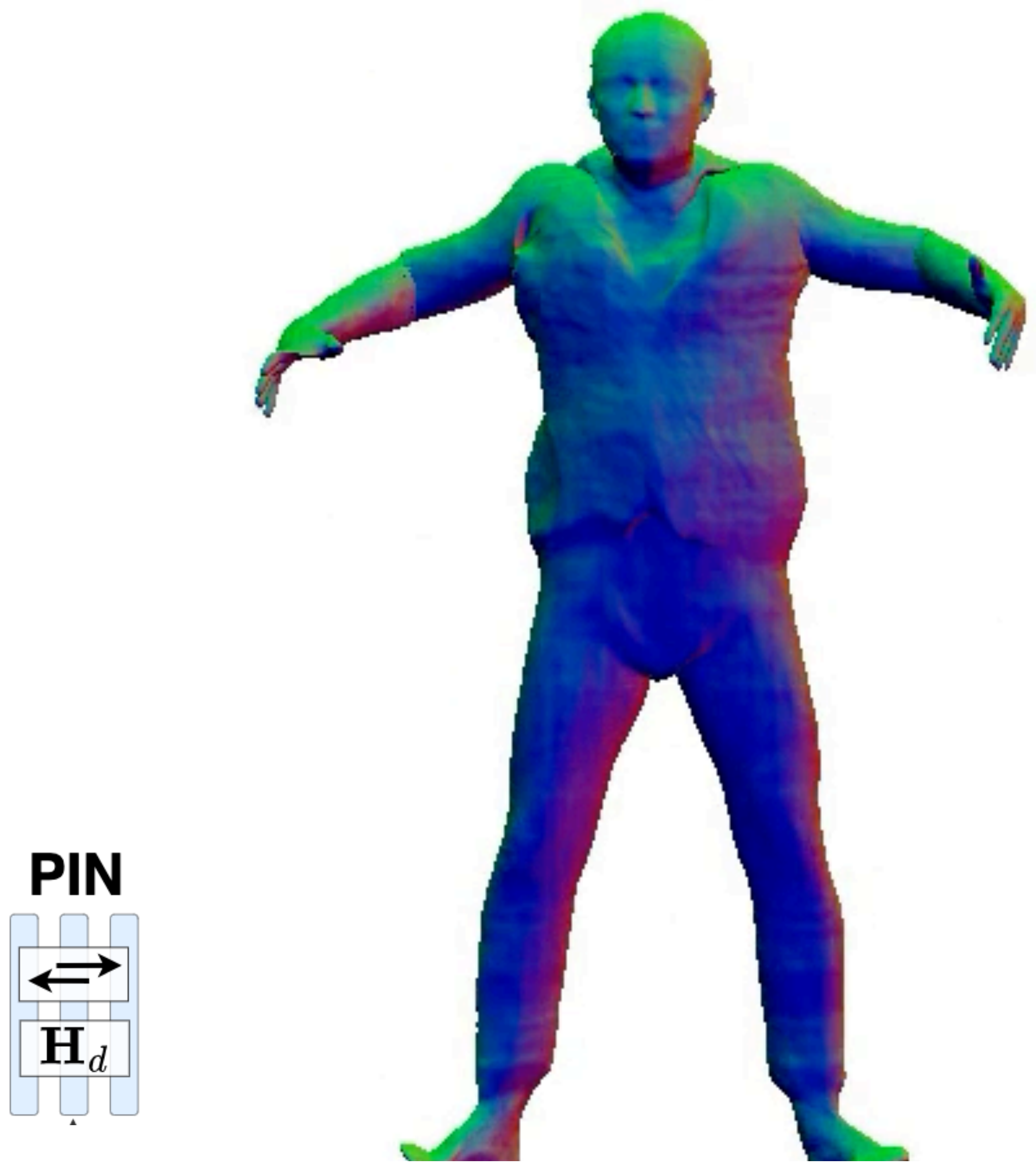
Pose Free Canonical Space



Pose deformations by $\mathbf{H}_c$



Target / Animation Pose



Pose deformations by $\mathbf{H}_d$

Ground Truth

INS output

Invertible Neural Skinning (Summary)

- an end-to-end **learnable** reposing technique**,**
- **preserves correspondences** across poses**,**
- more **accurate** and captures **pose-varying effects,**
- an **order of magnitude faster** than state-of-the-art.

Thanks!

Visit our poster on Thursday morning at CVPR.