Invertible Neural Skinning

Snap Inc.

Yash Kant[1,2] @yash2kant
Aliaksandr Siarohin[1] @NoTwitterYet
Rıza Alp Güler[1] @NoTwitterYet
Menglei Chai[1] @NoTwitterYet
Jian Ren[1] @NoTwitterYet
Sergey Tulyakov[1] @SergeyTulyakov
Igor Gilitschenski[2] @igilitschenski

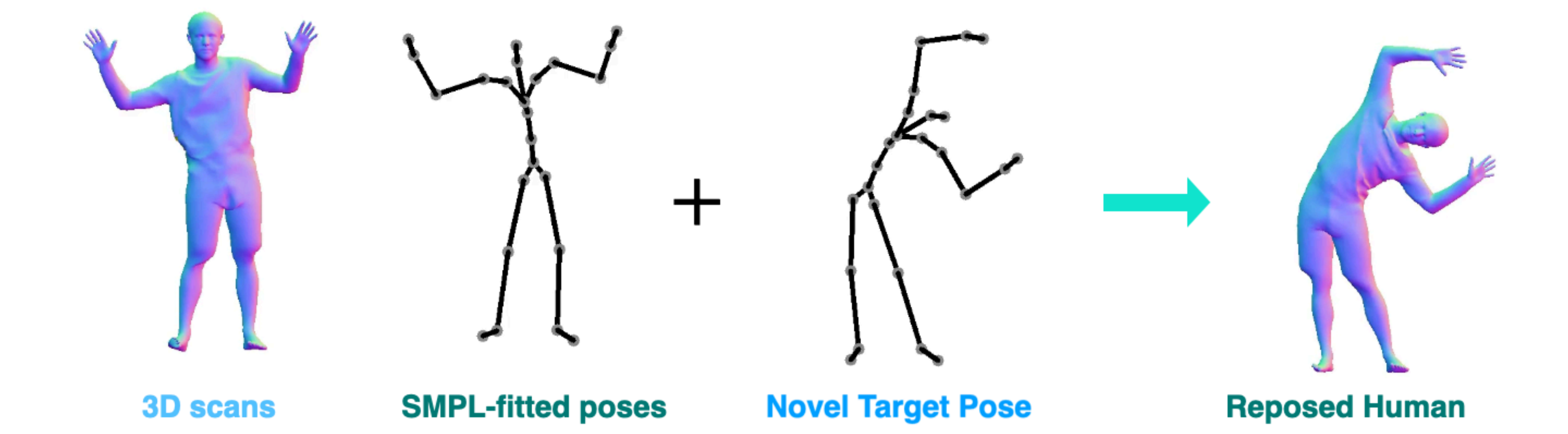
Snap Inc[1] University of Toronto[2]

Website

University of Toronto
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Overview

NVS Task: Given a single image of an object, generate novel views.



Task overview. Given a sequence of 3D scans and SMPL-fitted poses, we learn a reposing model.

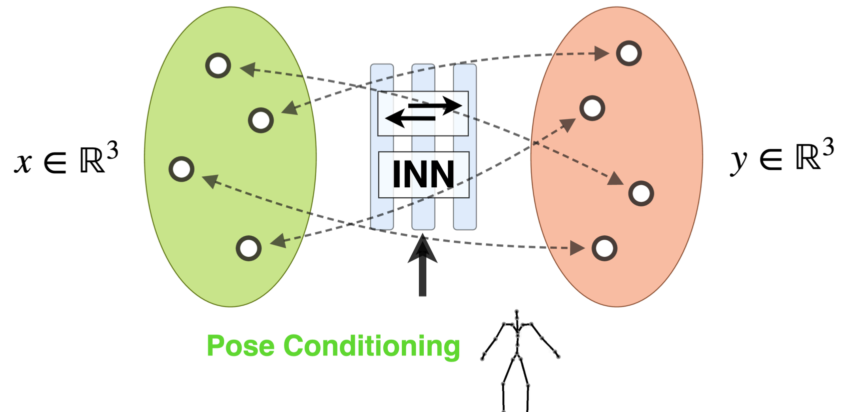
Issues with Existing Approaches.

- Limited expressivity of Linear Blend Skinning (LBS).
- Require costly mesh extraction (marching cubes) to generate each new pose.
- Correspondences not preserved across poses.

Contributions: End-to-end **learnable** setup, **preserves correspondences** across poses, more **accurate** and **order of magnitude faster** than state-of-the-art.

Background — LBS and Invertible Neural Network

Invertible Neural Network (INN) defines a bijective mapping between its input and output spaces, which we condition on relative body pose to create **Pose-driven INN (PIN).**



Pose-driven Invertible Neural Network (PIN)

Reposing task involves two main challenges.

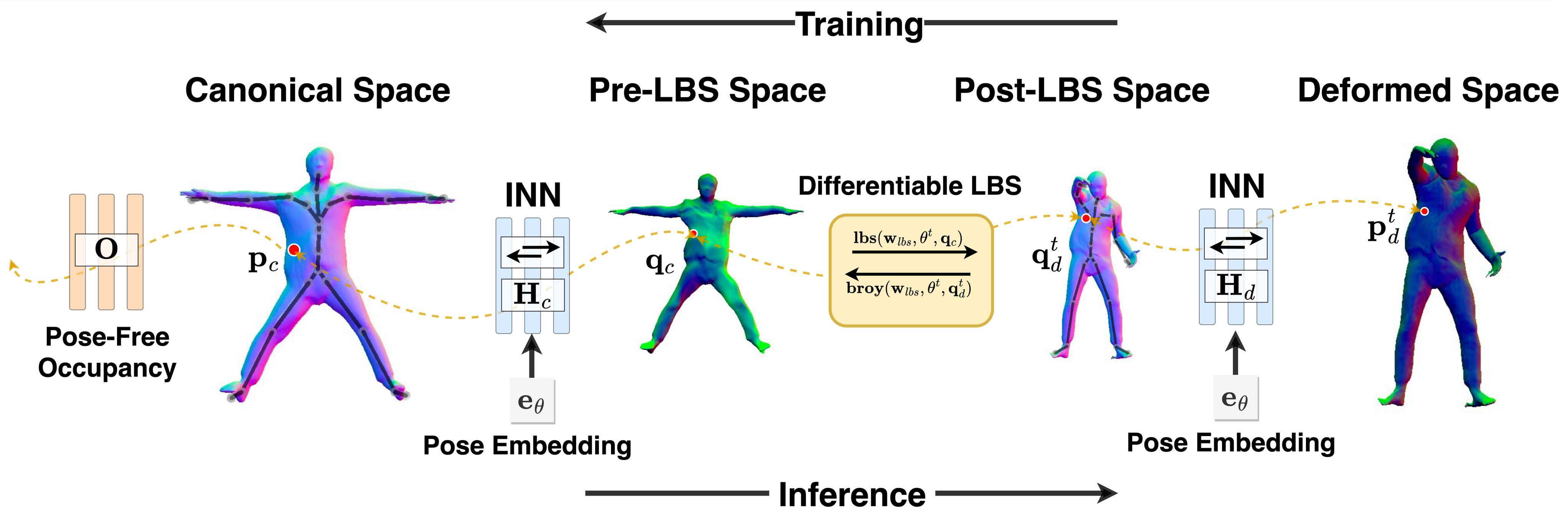
- Modeling the **movement of human body**, and
- Handling the **deformations that occur in clothing** due to body movement.

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

Linear Blend Skinning. 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.

Clothing Deformations. Linear Blend Skinning (LBS) **cannot capture non-linear deformations** of clothes and body tissue — as it interpolates linearly between poses.

Method



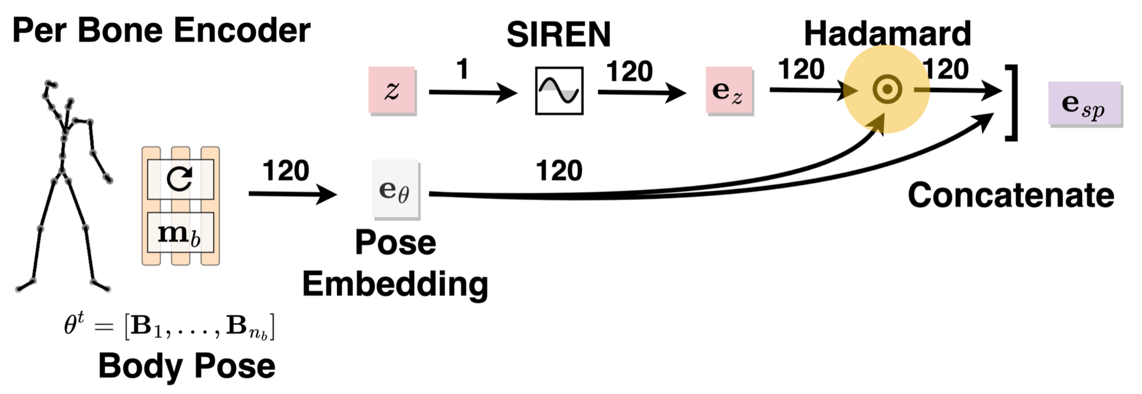
Invertible Neural Skinning. We chain **two PINs** around an **LBS** block to build final reposing pipeline.

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

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

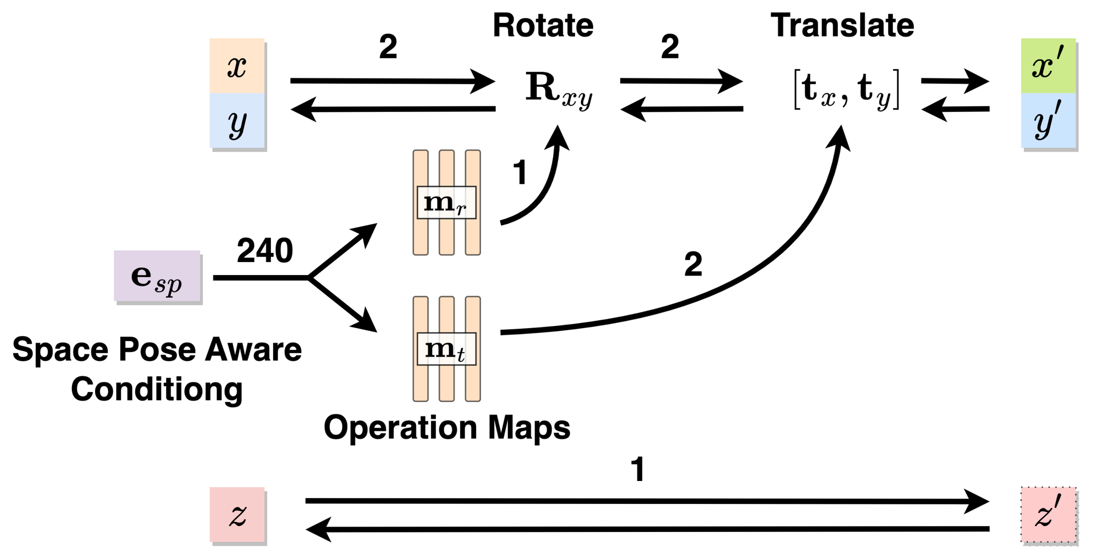
Efficient Reposing. Prior works (eg. SNARF) conditions canonical representation on target pose — which introduces non-linear deformations at the prices of expensive reposing (mesh extraction for each frame).

Bone Pose Encoder and Coupling Blocks



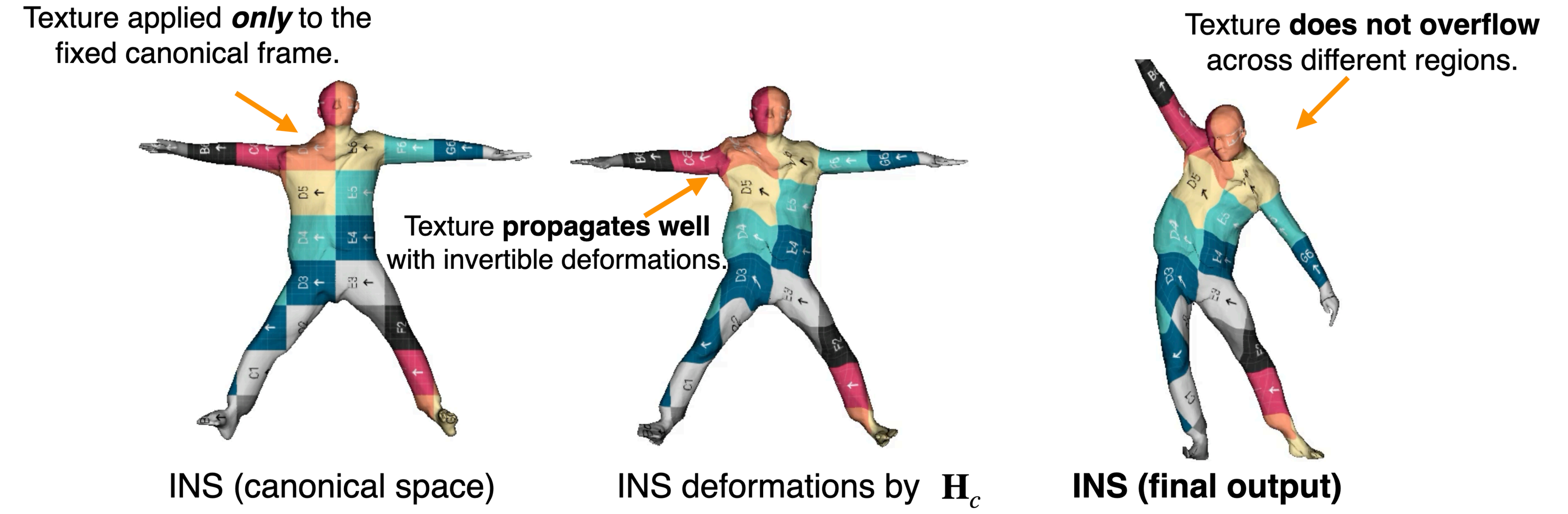
Space Pose Embedding. We encode each SMPL bone using an MLP, and combine it with the spatial part using scalar product.

Coupling Layer. Invertible networks are composed of coupling layers, which reversibly transform one part of the input (3D point) based on the other part as well as a provided condition (body pose).

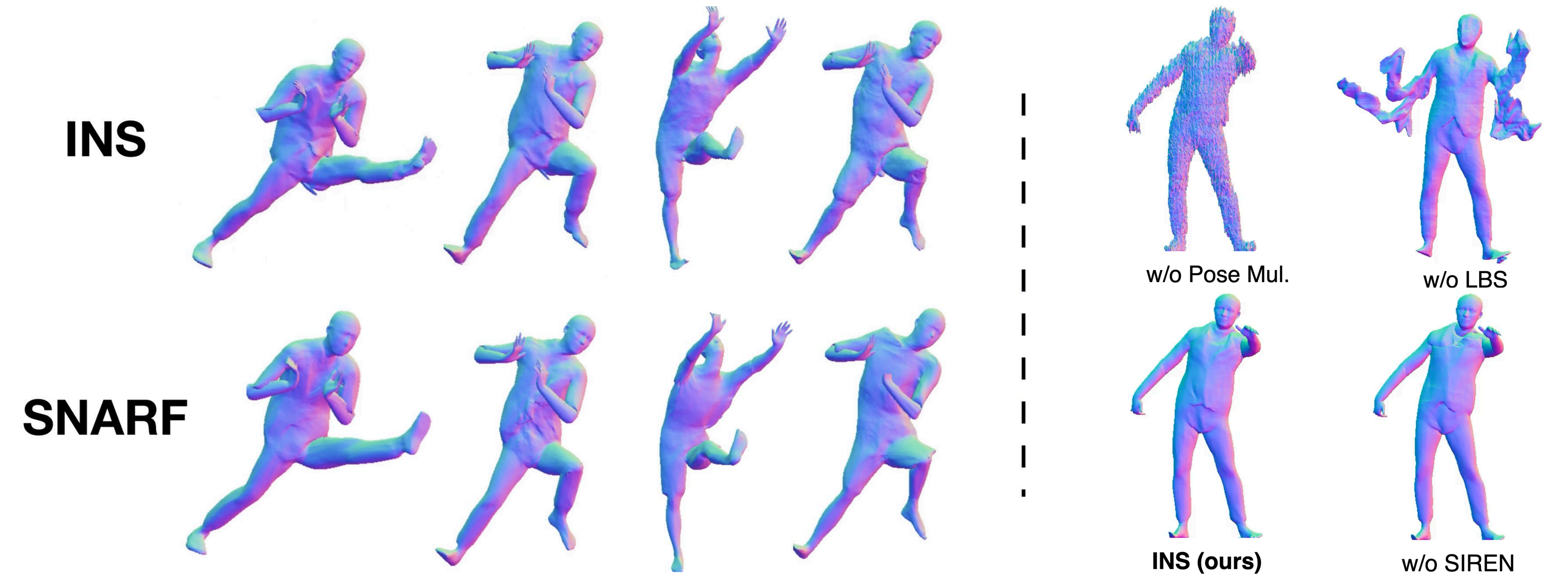


Results

Texture Propagation. As INS maintains correspondences between poses, it can propagate vertex attributes such as texture across poses.



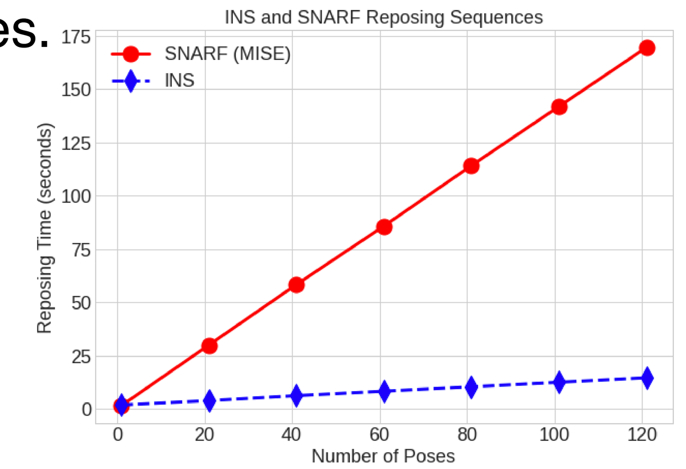
PosePrior. We find that INS produces more realistic cloth deformations compared to SNARF, while preserving limb extremities (fingers or feet) .



PosePrior Comparison between SNARF and INS. **Ablations in INS.**

Ablations. We find our pose multiplication strategy of combining pose and space to be crucial, as well as the use of SIREN and LBS modules.

Timing Study. Reposing the extracted mesh INS takes 0.13 seconds for an inference pass, which is an order of magnitude faster than 1.5 seconds taken by SNARF.



Metrics. We report the mean Intersection-over-Union of points sampled near the mesh surface (IoU Surface), and of points sampled uniformly in space (IoU Bounding Box).

		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%**

We match/outperform baselines on all metrics on CAPE (clothed human) dataset.