

Invertible Neural Skinning

Yash Kant^{1,2}, Aliaksandr Siarohin², Riza Alp Guler², Menglei Chai², Jian Ren², Sergey Tulyakov², Igor Gilitschenski¹

¹University of Toronto, ²Snap Research

<https://yashkant.github.io/invertible-neural-skinning/>

Abstract

Building animatable and editable models of clothed humans from raw 3D scans and poses is a challenging problem. Existing reposing methods suffer from the limited expressiveness of Linear Blend Skinning (LBS), require costly mesh extraction to generate each new pose, and typically do not preserve surface correspondences across different poses. In this work, we introduce Invertible Neural Skinning (INS) to address these shortcomings. To maintain correspondences, we propose a Pose-conditioned Invertible Network (PIN) architecture, which extends the LBS process by learning additional pose-varying deformations. Next, we combine PIN with a differentiable LBS module to build an expressive and end-to-end Invertible Neural Skinning (INS) pipeline. We demonstrate the strong performance of our method by outperforming the state-of-the-art reposing techniques on clothed humans and preserving surface correspondences, while being an order of magnitude faster. We also perform an ablation study, which shows the usefulness of our pose-conditioning formulation, and our qualitative results display that INS can rectify artefacts introduced by LBS well.

1. Introduction

Being able to create animatable representations of clothed humans beyond skinned meshes is essential for building realistic augmented or virtual reality experiences and improving simulators. Towards this goal, we consider the task of building animatable human representations from raw 3D scans and corresponding poses. Prior work in this area has seen a shift from building parametric models of humans [6, 22, 31], to more recent works learning implicit 3D neural representations [1, 12, 13, 50, 51, 55, 56] from data in canonical space. These canonical representations are animated to a new pose by a learning skinning weight field around them [11, 14, 37, 52, 57] and applying Linear Blend Skinning (LBS) to warp the surface, where the pose is defined by a bone skeleton underlying the 3D surface.

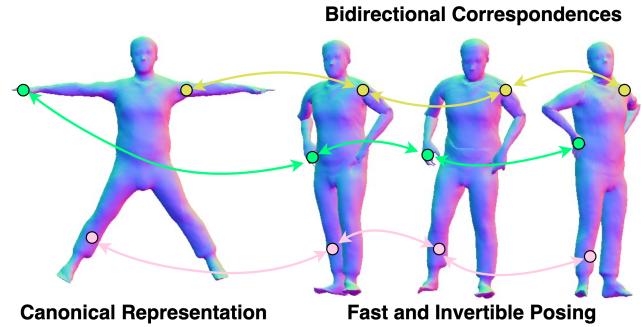


Figure 1. **Fast and Invertible Posing.** We propose an end-to-end learnable reposing pipeline that allows animating implicit surfaces with intricate pose-varying effects, without requiring mesh extraction [36] for each pose, while also maintaining correspondences across poses.

These prior works generally suffer from the limited expressivity of LBS when handling complex pose-varying deformations, such as those of loose clothes and body tissue (i.e. muscle bulges, skin wrinkles). In parametric models like SMPL [31], such deformations are represented by adding simple linear pose correctives (aka blend shapes), but these are restrictive and only work for unclothed humans. Implicit methods, to relieve this issue, learn their canonical representations conditioned on the deformed pose [11, 14]. However, this conditioning comes with two major drawbacks during reposing. Given the sequence of poses, a new mesh has to be extracted from scratch for each pose, which becomes a bottleneck when animating subjects at a high frame-rate or resolution. Also, as a consequence of this step, correspondences (topology preservation) between the surfaces of the same subject across different poses are lost.

Invertible Neural Networks (INN) [15, 16, 26] are bijective functions that can preserve exact correspondences between their input and output spaces, while learning complex non-linear transforms between them. This ability of INNs makes them a suitable candidate for reposing, and in this work, we leverage INNs to build an Invertible Neu-

ral Skinning (INS) pipeline. For this, we first build a Pose-conditioned Invertible Network, abbreviated as PIN¹, to learn pose-conditioned deformations. Next, to create an end-to-end Invertible Neural Skinning (INS) pipeline, we place two PINs around a differentiable LBS module, and use a pose-free canonical representation. These PINs help capture the non-linear surface deformations of clothes across poses and alleviate the volume loss suffered from the LBS operation. Since our canonical representation remains pose-free, we perform the expensive mesh extraction exactly once, and repose the mesh by simply warping it with the learned LBS and an inverse pass through PINs.

We demonstrate the strong performance of INS by outperforming the previous state-of-the-art reposing method SNARF [11]. On clothed humans data, we find INS provides an absolute gain of roughly 1% when compared to SNARF with pose-conditioning, and roughly 6% compared to SNARF without pose-conditioning. We conduct experiments on much simpler minimally clothed human data and obtain competitive results. We also find INS to be an order of magnitude faster at reposing long sequences. We ablate our INS and demonstrate the effectiveness of our pose-conditioning formulation. Our results clearly show that the proposed INS can correct the LBS artefacts well.

2. Related Work

Representing Articulate Characters in 3D. Over the years, a significant amount of prior work for building parametric representations of the human body [2, 22, 30, 31, 40, 45, 61] or for specific parts such as hands and faces [39, 49] was developed. Beyond humans, recent work developed parametric animal models [4, 5, 65]. Encouraged by the rapid progress in implicit neural 3D representations [36, 38], a number of works explored building implicit human representations with and without clothing [20, 23, 33, 53, 57–60, 63, 64]. Representing characters as implicit functions comes with a cost of time-consuming mesh extraction via Marching Cubes [32].

Animating 3D Representations with Poses. Parametric models usually define the correspondences between poses, represented as a set of bones, and mesh vertices through Linear Blend Skinning (LBS) weights. These weights provide a soft assignment of vertices to human bones. Thus for animation, these models simply transform the vertices using a linear combination of bone transformations. When the parametric model is not available, these weights need to be discovered. To this end, recent works adopt learning-based solutions for discovering LBS weights [10, 11, 14, 29, 37, 46, 47, 52]. They usually assume a shared canonical space and learn a canonical LBS weight field, which

¹We avoid the abbreviation PINN to avoid confusion with Physics-inspired Neural Networks [48]

is used for deforming the body in the novel pose during inference. However, at training time, the character needs to be warped backward from deformed to canonical space, i.e. given deformed points, we need to obtain corresponding canonical points. Thus some works [8, 17, 52, 56, 62] learn LBS weights separately in deformed and canonical spaces, which could be used for establishing correspondences. These generally require cycle-consistency losses for regularization. Recently, SNARF [11] proposed to compute these correspondences by finding the solutions of the LBS equation using an iterative solver. We adopt a similar formulation, to discover the correspondences as well.

However, LBS is often insufficient to capture non-linear deformations of flowy clothes and body tissue (i.e. muscle bulges). To mitigate this problem, prior works [11, 14, 58] condition their canonical representations on the deformed pose. Such conditioning helps to alleviate the shortcomings of pure-LBS deformations, but this comes at a cost following two major limitations:

- **Slow Reposing.** To generate a new animation given a sequence of poses across time, these methods extract a separate mesh from scratch for each pose. This becomes a bottleneck if we want to pose the character at a high frame-rate or resolution.
- **No Correspondences.** As a consequence of the above step, two completely separate meshes get extracted at each pose with no correspondence between them.

In our work, we address both these limitations by extending pure-LBS formulation with additional Pose-conditioned Invertible Networks (PIN), while using a pose-free canonical representation.

Invertible Neural Networks for 3D Vision. INNs [15, 16, 18, 21, 24, 26, 41] were initially designed for tractable density estimation of high-dimensional and generative modeling, a.k.a. Normalizing Flows [3, 27]. Usually, INNs are built by chaining together multiple conditional *Coupling Layers* [15, 16], where a single coupling layer defines an invertible transformation between its input and output. The main idea behind *Coupling Layers* is that if we split the input into two parts and only modify the first part while conditioning this modification on the second, this should be trivially invertible. Another popular type of invertible transformations are *Invertible Residual layers* [24] with small conditioning numbers. They utilize fixed point iterations for finding an inverse. In our work, we mostly rely on *Coupling Layers* since they are faster, and we did not see any additional benefits from *Residual layers*. In the context of 3D vision, INNs were explored for learning primitives of 3D representations [43], doing 3D shape-completion tasks [28], and reconstructing dynamic scenes [9]. However, to the best

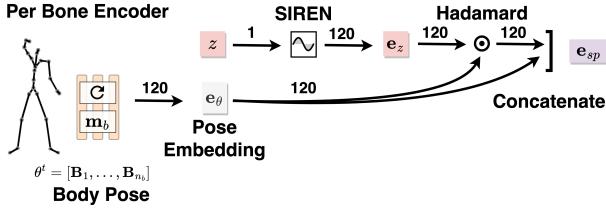


Figure 2. **Space and Pose Aware Conditioning.** We encode the body pose using a per-bone MLP network operating on individual bone transforms. The pose embedding is then fused with space embedding to generate conditioning for PIN.

of our knowledge, their usage has not been explored for animating 3D characters.

3. Method

Task Setup. The goal of our work is to learn a human 3D representation that allows the generation of novel poses beyond original training data (a.k.a. reposing). For each subject, we assume the availability of N pairs consisting of bone poses and 3d meshes denoted as $(\theta^t, \mathbf{M}^t)_{t=1}^N$. Such data can be obtained from human scans, and the poses can be estimated by fitting a parametric SMPL-like body model to these scans. *Given this data, we wish to learn a subject-specific implicit neural representation in a canonical space and a method to animate this representation.*

Deformed and Canonical Spaces. We denote an input point in deformed space as $\mathbf{p}_d^t \in \mathbb{R}^3$ and a point in the canonical space as $\mathbf{p}_c \in \mathbb{R}^3$. Since our input consists of a sequence of deformed (posed) meshes, we use the superscript t to indicate the time-step of capture. As our canonical space is independent of the pose, it is shared across all the time steps; hence, \mathbf{p}_c is not time-indexed.

Deformed and Canonical Poses. We follow the SMPL model [31], which represents body pose as a set of bones in a kinematic tree. While reposing, as we only require the relative pose between canonical and deformed space at any given time t , we represent this by $\theta^t = [\mathbf{B}_1, \dots, \mathbf{B}_{n_b}]$, where $\mathbf{B}_i = [\mathbf{R}_i | \mathbf{t}_i]$ represents a transformation of the i^{th} bone in 3D space, i.e. $\mathbf{B}_i \in \text{SE}(3)$ with corresponding rotation $\mathbf{R}_i \in \mathbb{R}^{3 \times 3}$ and translation $\mathbf{t}_i \in \mathbb{R}^3$. We denote the total number of bones by n_b .

Pose-free Canonical Occupancy. To represent a specific subject, we use an Occupancy Network \mathbf{O} [36] conditioned solely on the input point \mathbf{p}_c . The canonical surface S_c is then represented implicitly as a level-set ($\sigma = 0.5$) of this occupancy network.

$$S_c = \{\mathbf{p}_c \mid \mathbf{O}(\mathbf{p}_c) = \sigma\} \text{ and } \mathbf{O} : \mathbb{R}^3 \rightarrow [0, 1]. \quad (1)$$

To extract this canonical iso-surface as a mesh, we use the MISE [36] algorithm. This is different from previous

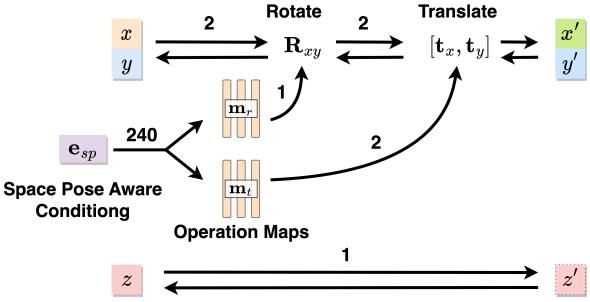


Figure 3. **Pose-conditioned 2D Coupling Layer.** We use the space-pose conditioning to predict the operation parameters using two operation maps (MLPs), and use them to rotate and translate the input split $[x, y]$. In this case, $[z]$ remains unchanged.

works [11, 14] that use additional pose-conditioning in the canonical occupancy network.

Sampling Points. For both training and evaluation of INS, we sample 3D points in deformed space and get their ground-truth occupancy values of zero or one based on whether they lie outside or the mesh (scan). We put exact details on this sampling in Appendix A.1.

3.1. Differentiable Forward Blend Skinning

To animate our subject from their canonical to deformed pose we use Linear Blend Skinning (LBS), which involves deforming the canonical surface according to a convex combination of rigid bone transforms. Specifically, we use the differentiable LBS formulation from SNARF [11] and summarize it below.

Canonical Weight Field. We define a learnable weight field in canonical space parameterized by a neural network, $\mathbf{w}_{lbs} : \mathbb{R}^3 \rightarrow \mathbb{R}^{n_b}$. For a given point in canonical space, this weight field predicts the blend weights corresponding to each bone:

$$\mathbf{w}_{lbs}(\mathbf{q}_c) = [w_1, \dots, w_{n_b}] \text{ and } w_i \in \mathbb{R}. \quad (2)$$

To make weights (w_i) convex for LBS, they are constrained to be always non-negative and sum to 1 using softmax.

LBS. Given the above weight field and the relative body pose as bone transforms $\theta^t = [\mathbf{B}_1, \dots, \mathbf{B}_{n_b}]$, we can forward warp any point \mathbf{q}_c of our canonical space to deformed space using Linear Blend Skinning as follows:

$$\mathbf{q}_d^t = \text{lbs}(\mathbf{w}_{lbs}, \mathbf{q}_c, \theta^t) = \left[\sum_{i=1}^{n_b} \mathbf{w}_{lbs,i}(\mathbf{q}_c) \cdot \mathbf{B}_i \right] \cdot \mathbf{q}_c \quad (3)$$

where \mathbf{q}_d^t represents the corresponding point in deformed space where \mathbf{q}_c lands after LBS.

Searching Canonical Correspondences. While training on raw scans, we are only provided with points in deformed space. To find their possible correspondences

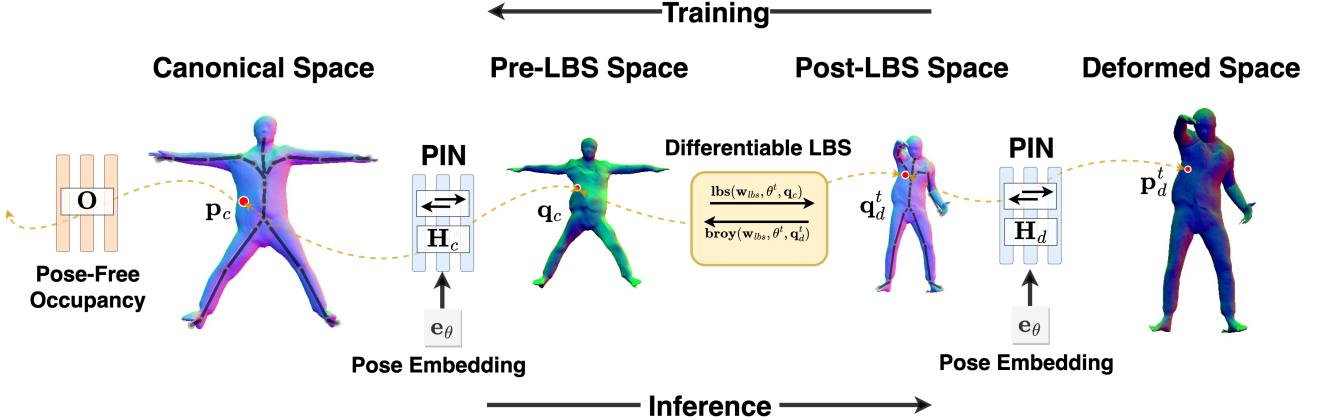


Figure 4. **Invertible Neural Skinning.** Our end-to-end differentiable reposing pipeline consists of two Pose-conditioned Invertible Networks (PINs) placed around a differentiable LBS block. These PINs (\mathbf{H}_c and \mathbf{H}_d) capture non-linear surface deformations of clothes and attenuate LBS artefacts. Our canonical representation is not conditioned on the target pose and requires mesh extraction only once. The green shade in Pre-LBS and Deformed spaces indicate the deformations introduced by PINs, and its intensity denotes their magnitudes.

in canonical space, we solve for the roots of Equation 3 using an iterative solver while keeping \mathbf{w}_{lbs} constant. Specifically, we use Broyden’s Method [7] to find a set of $\{\mathbf{q}_c^1, \dots, \mathbf{q}_c^K\}$ point correspondences for each deformed point \mathbf{q}_d^t by initializing the root-finding algorithm at K different points in the canonical space.

$$\{\mathbf{q}_c^1, \dots, \mathbf{q}_c^K\} = \text{broy}(\mathbf{w}_{lbs}, \theta^t, \mathbf{q}_d^t). \quad (4)$$

Differentiable Skinning. The above formulation is end-to-end differentiable as it is possible to compute the gradients of the weight field \mathbf{w}_{lbs} with respect to input point \mathbf{q}_d^t via implicit differentiation as shown in SNARF [11], Section 3.4. In this work, we also extend these derivations to compute the gradient of correspondences \mathbf{q}_c^i w.r.t. input points. For more details, please refer to Appendix D.

Why is LBS insufficient? The above differentiable formulation suffers from the same limitations of traditional LBS, such as being unable to represent the clothed surfaces, and introducing volume loss, as shown in Figure 6, b). This is especially problematic when learning from real-world data of clothed humans in various poses.

3.2. Pose-conditioned Invertible Network (PIN)

Invertible networks [15, 16, 26] are bijective functions composed of modular components called coupling layers, which preserve 1-1 correspondences between their input and output. In this section, we describe the construction of our proposed pose-conditioned coupling layer, which is chained together to construct a PIN.

2D Coupling Layer (Figure 3). A coupling layer operates by splitting its input into two parts using a fixed breaking pattern. After splitting, the first part of the input is transformed by applying a sequence of invertible operations, such as translation and rotation. The parameters for

these operations can be produced by any arbitrary function that is jointly conditioned on the second part of the input and an external conditioning, such as pose in our case.

Formally, as we operate in 3D space, let the input point be defined as $[x, y, z]$, and the input splits are $[x, y]$ and $[z]$. Then the 2D coupling layer $\mathbf{G}^{xy}([x, y, z], \theta^t)$ defines an invertible transformation as follows:

$$[x', y'] = \mathbf{R}_{xy}[x, y]^T + [t_x, t_y] \quad \text{and} \quad z' = z, \quad (5)$$

where $\mathbf{R}_{xy} \in \mathbb{R}^{2 \times 2}$, and $[t_x, t_y] \in \mathbb{R}^2$ is a rotation matrix and translation vector produced by any arbitrary function that takes as input only the bone pose θ^t and the coordinate z . The inverse $\mathbf{G}_{xy}^{-1}([x, y, z], \theta^t)$ of the coupling layer can be computed easily by:

$$[x, y] = \mathbf{R}_{xy}^{-1}([x', y'] - [t_x, t_y]) \quad \text{and} \quad z = z'. \quad (6)$$

We describe computation of operation parameters \mathbf{R}_{xy} and $[t_x, t_y]$ next.

Pose Embedding. We encode every bone transform in pose θ^t using a MLP \mathbf{m}_b which takes a 6D input of concatenated 3D translation and rotation (as Euler angles). To obtain pose embedding, we concatenate the outputs of each bone \mathbf{e}_θ as follows:

$$\mathbf{m}_b : \mathbb{R}^6 \rightarrow \mathbb{R}^{d/n_b} \quad \text{and} \quad \mathbf{e}_\theta := \text{concat}[\mathbf{m}_b(\mathbf{B}_i)]_{i=1}^{n_b}. \quad (7)$$

Space Embedding. We use SIREN [54], a learned and periodic positional encoding, to map the spatial coordinates denoting it as

$$\mathbf{e}_z := \Phi(z) : \mathbb{R}^1 \rightarrow \mathbb{R}^d. \quad (8)$$

We find that this helps to better represent high-frequency surface details such as cloth wrinkles.

Space and Pose Aware Conditioning (Figure 2). We observe that when the relative pose θ^t between deformed and canonical spaces is zero (i.e. $B_i = [\mathbf{I}|0]$, all bone transforms have identity rotation and zero translation), the coupling layer should not introduce any space-varying (i.e. z -conditioned) changes.

To enforce this, we take the Hadamard product of the space and pose embeddings, and subsequently concatenate them obtaining

$$\mathbf{e}_{sp} := \text{concat}[\mathbf{e}_\theta \odot \mathbf{e}_z, \mathbf{e}_\theta] \in \mathbb{R}^{2d}. \quad (9)$$

We visualize the construction of our space and pose aware conditioning \mathbf{e}_{sp} in Figure 2.

Rotation and Translation Maps. Finally, to produce parameters for coupling operations, we use two MLPs \mathbf{m}_t and \mathbf{m}_r , which take as input the above conditioning vector:

$$[t_x, t_y] = \mathbf{m}_t(\mathbf{e}_{sp}) : \mathbb{R}^{2d} \rightarrow \mathbb{R}^2, \quad (10)$$

$$\gamma_{xy} = \mathbf{m}_r(\mathbf{e}_{sp}) : \mathbb{R}^{2d} \rightarrow \mathbb{R}^1. \quad (11)$$

Note that the output of \mathbf{m}_r only predicts the angle of rotation γ_{xy} in radians (a single value). The axis of rotation passes through the origin of the split input space, i.e. XY space here. We convert γ_{xy} into a rotation matrix \mathbf{R}_{xy} .

1D Coupling Layer. Unlike the 2D coupling layer described above, we cannot use the rotation operator in 1D, and in this case, we only use translation. For layer $\mathbf{G}^x([x, y, z], \theta^t)$ with split pattern as $[x]$ and $[y, z]$ the coupling operation becomes:

$$x' = x + t_x \text{ and } [y', z'] = [y, z], \quad (12)$$

where $t_x \in \mathbb{R}^1$ produced in a similar fashion as a 2D coupling layer using a translation map $\mathbf{m}_t : \mathbb{R}^{2d} \rightarrow \mathbb{R}^1$ with single scalar output instead of 2D translation. The space embedding of Equation 8, in the 1D case, takes both coordinates as input $\mathbf{e}_{xy} = \Phi([x, y]) : \mathbb{R}^2 \rightarrow \mathbb{R}^d$.

Pose-conditioned Invertible Network (PIN). Finally, we compose our PIN by chaining together multiple 1D and 2D pose-conditioned coupling layers as:

$$\mathbf{H}(\mathbf{p}, \theta^t) = \mathbf{G}_1 \circ \mathbf{G}_2 \circ \dots \mathbf{G}_n : \mathbb{R}^3 \rightarrow \mathbb{R}^3, \quad (13)$$

where \mathbf{p} represents point in 3D space, \mathbf{G}_i represents a coupling layer, and θ^t represents pose. Inverting PIN is simply equivalent to sequentially inverting each coupling layer in the reverse order:

$$\mathbf{H}^{-1}(\mathbf{p}, \theta^t) = \mathbf{G}_n^{-1} \dots \mathbf{G}_2^{-1} \circ \mathbf{G}_1^{-1}. \quad (14)$$

Since the PIN is invertible by construction, it preserves exact correspondences between its input and output spaces:

$$\mathbf{p} = \mathbf{H}^{-1}(\mathbf{H}(\mathbf{p}, \theta^t), \theta^t) \quad \forall \mathbf{p} \in \mathbb{R}^3. \quad (15)$$

We visualize a single coupling layer of PIN in Figure 3.

3.3. Invertible Neural Skinning

Overview (Figure 4). Our overall posing pipeline INS is comprised of three previously described components chained together:

- \mathbf{H}_c : A Pose-conditioned Invertible Network (PIN) \mathbf{H}_c that operates after canonical space and before LBS.
- \mathbf{H}_d : A Pose-conditioned Invertible Network (PIN) that operates before deformed space and after LBS.
- Differentiable LBS network as described in Section 3.1 operating between above PINs.

Next, we discuss how we formulate an invertible mapping that preserves correspondences between deformed and canonical spaces.

Deformed to Canonical (Training). For any point \mathbf{p}_d^t in deformed space, we first process it using PIN \mathbf{H}_d , and obtain \mathbf{q}_d^t . Next, we use Broyden’s algorithm to get correspondences of \mathbf{q}_d^t in canonical space, let’s say $\{\mathbf{q}_c^i\}_{i=1}^K$. Finally we use a second PIN \mathbf{H}_c to map these points $\{\mathbf{p}_c^i\}_{i=1}^K$ in the pose-independent canonical space.

$$\mathbf{p}_d^t \xrightarrow{\mathbf{H}_d(\cdot, \theta^t)} \mathbf{q}_d^t \xrightarrow{\text{broy}(\cdot, \theta^t)} \{\mathbf{q}_c^i\} \xrightarrow{\mathbf{H}_c(\cdot, \theta^t)} \{\mathbf{p}_c^i\}. \quad (16)$$

To obtain the most suitable canonical correspondence, we take the arg max over all predicted canonical occupancies

$$\mathbf{p}_c^* = \arg \max_{i=1 \dots K} \{\mathbf{O}(\mathbf{p}_c^i)\}. \quad (17)$$

During training, we approximate the arg max with a softmax function in order to backpropagate gradients softly through all correspondences following SNARF.

Training Objective. In our dataset, we are given points in deformed space and corresponding ground truth occupancy values of zero or one. We map these deformed points to canonical space and apply binary cross-entropy loss to jointly train all components of the posing network according to

$$\min_{\mathbf{H}_d, \mathbf{H}_c, \mathbf{w}_{lbs}, \mathbf{O}} \mathcal{L}_{bce}(\mathbf{O}(\mathbf{p}_c), o_{gt}). \quad (18)$$

Auxiliary Objectives. Following SNARF, we enforce a prior on the canonical pose by using two additional losses during the first epoch. First, we sample additional points on bones in canonical pose and encourage their occupancies to be one. Second, we encourage the skinning weight of bone joints to be equal. However, no ground truth skinning weights are required during these steps.

Canonical to Deformed (Inference). Once trained, we can animate characters using INS in any given novel pose θ^n in two simple steps. First, running mesh extraction on the canonical occupancy network. Second, reposing the mesh vertices via an inverse pass of our posing pipeline as follow:

$$\mathbf{p}_c \xrightarrow{\mathbf{H}_c^{-1}(\cdot, \theta^n)} \mathbf{q}_c \xrightarrow{\text{lbs}(\cdot, \theta^n)} \mathbf{q}_d^t \xrightarrow{\mathbf{H}_d^{-1}(\cdot, \theta^n)} \mathbf{p}_d^t. \quad (19)$$

Subject	Clothing	IoU Surface					IoU Bounding Box				
		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%

Table 1. **Quantitative Results on Clothed Humans.** We find our approach INS outperforms all methods when averaged across 15 runs, on both IoU Surface and IoU Bounding Box metrics.

Subject	IoU Surface			IoU Bounding Box		
	SNARF	SNARF-NC	INS (ours)	SNARF	SNARF-NC	INS (ours)
Average	90.01%	85.22%	88.59%	97.21%	95.72%	96.35%

Table 2. **Quantitative Results on Minimally Clothed Humans.** On DFAUST, INS outperforms SNARF-NC by a large margin while performing competitively with SNARF, and being order magnitude faster at reposing.

Fast Reposing. As our canonical occupancy network \mathbf{O} is independent of θ^m we only have to extract mesh exactly once. And reposing this mesh for a sequence of poses simply becomes equivalent to performing multiple inferences described in Equation 19.

4. Experiments

4.1. Evaluation

Datasets. Training our method requires sampled points in the deformed space, along with corresponding occupancies and poses. Thus, we benchmark INS on two datasets CAPE [34], which features scanned humans in loose clothing, and DFAUST [35], containing only minimally clothed human scans.

CAPE [34] contains scans of 15 subjects (8 males and 7 females), wearing 8 different types of garments while performing a large number of actions. These actions were recorded using a high-resolution body scanner (3dMD LLC, Atlanta, GA), and the scans were registered using an SMPL model [31]. Similarly to SNARF [11], INS requires training a new model every *subject-cloth* pair, and exhaustively training on every combination quickly becomes expensive. To manage computational costs, we use a subset of 15 sequences. This subset covers all garment types at least once and most of the subjects, thus capturing variations in both — body shape and clothing.

DFAUST is a subset of the AMASS [35] dataset consisting of 10 subjects who are minimally clothed. Each subject is scanned similarly to CAPE while performing 10 different actions. As the subjects wear minimal clothing, much of their motions can be represented accurately by rigid body transformations. As a consequence of little subject clothing, we observe that DFAUST contains significantly fewer pose-specific deformations. Thus we note that DFAUST dataset is not well-posed to test the true capabilities of INS.

Data Splits. For a given subject in DFAUST or a subject-clothing pair in CAPE, we are provided with multiple temporal sequences, each containing a different action. We di-

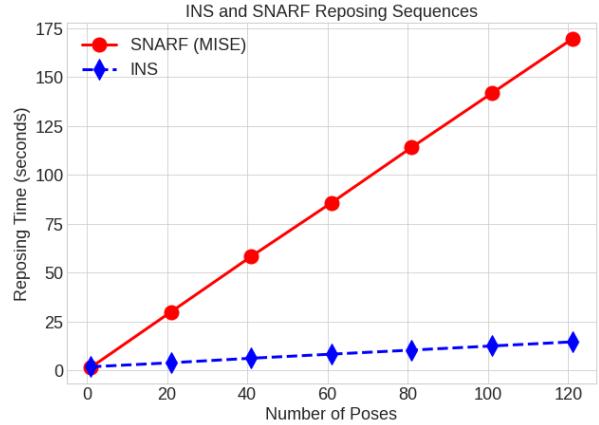


Figure 5. **Reposing time comparison between INS and SNARF**
We show the time taken by SNARF vs INS for reposing a mesh extracted at 128^3 resolution across 125 different target poses. INS performs reposing an order of magnitude faster than SNARF.

vide these sequences in 9:1 ratio into train and test sets. This split is similar to SNARF [11]. More details about training sequences and garment types are provided in Appendix B.

Metrics. Following SNARF [11], we report the mean Intersection-over-Union of points sampled near the mesh surface (IoU surface), and of points sampled uniformly in space (IoU bbox).

4.2. Baselines

SNARF-NC. We use SNARF [11] without pose-conditioning in the canonical occupancy network as our first and primary baseline. For this, we only remove the pose-conditioning used by SNARF such that canonical space becomes pose-independent, i.e. $\mathbf{O}(\mathbf{p}_c)$. We do not make any other changes. This setting is comparable to INS as it allows for fast posing and preserves correspondences across different poses.

SNARF. We also compare INS to the original SNARF [11] which uses a pose-conditioned occupancy network, i.e. $\mathbf{O}(\mathbf{p}_c, \theta^t)$. However, we point out that the above pose-conditioned occupancy comes at the sacrifice of fast posing, by requiring expensive mesh extraction for each new pose while not preserving correspondences across them. These disadvantages make the direct comparison between INS and SNARF based solely on their performance a little lopsided.

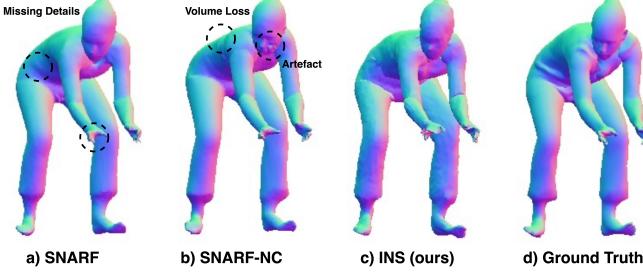


Figure 6. **Qualitative Samples on CAPE.** We find that SNARF (leftmost) struggles to represent finer details such as cloth wrinkles. Whereas SNARF-NC (second left) struggles with LBS artefacts such as volume loss, and candy-wrapper effects. Meanwhile INS (second right) is able to repose surfaces while capturing sharper local-details. Best viewed under zoom.

To obtain SNARF and SNARF-NC results, we use the official codebase² released by the authors.

AVG-LBS. In addition to the above strong learned baselines, we provide results on two simpler baselines, which use the SMPL-fitted LBS weights to unpose the meshes (scans) using forward skinning. For this, we simply take an average of all the canonicalized training meshes to generate a final canonical mesh and deform it to any unseen given pose using Foward LBS and SMPL weights.

FIRST-LBS. This baseline is similar to the AVG-LBS baseline described above and uses SMPL-fitted weight for reposing. Instead of using an average across all training meshes, it only uses the first mesh, thus containing lesser pose-conditioned details.

4.3. Main Results

CAPE. We demonstrate the results of INS on clothed human data in Table 1. Given the challenging nature of modeling cloth deformations contained in this dataset, we find that INS surpasses SNARF-NC (without pose-conditioning) on average by **+6.24%** and **+6.41%** absolute percentage points in Surface IoU and Bounding Box IoU respectively. Moreover, INS also outperforms vanilla SNARF with pose-conditioning by **+0.89%** and **+1.02%** absolute percentage points in Surface IoU and Bounding Box IoU, respectively, while also enjoying the benefits of fast posing, and matched correspondences across various poses. We observe that the simple aggregation baseline of AVG-LBS performs quite closely with SNARF-NC with a performance drop of only 1.88% and 1.66% percentage points between them. However, AVG-LBS benefits from using a strong prior of parametric SMPL model and corresponding fitted weights.

DFAUST. We also report results on much simpler minimally clothed humans from the DFAUST dataset in Table 2. We find that INS outperforms SNARF-NC (without pose-conditioning) on average by **+3.37%** and **+0.63%**

²<https://github.com/xuchen-ethz/snarf>

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

Table 3. **Ablation Table.** We perform an ablation study of INS on a clothed subject 03375 (Table 1, Row 1) from the CAPE dataset.

absolute percentage points in Surface IoU and Bounding Box IoU metrics, respectively. When compared to SNARF with pose conditioning, we find INS lags behind by **-1.42%** and **-0.86%** absolute percentage points in Surface IoU and Bounding Box IoU metrics, respectively. Given the minimal clothing and few pose-conditioned non-linear effects in DFAUST, we hypothesize that this performance drop can be attributed to SNARF overfitting easily to this benchmark. We believe this result also reflects the importance of testing on many realistic datasets such as CAPE.

Timing Study. In Figure 5, we compare the times taken by SNARF and INS to repose a clothed character across a sequence of 125 different poses. A single mesh extraction pass with MISE [36] operating on the cube of resolution 128^3 , takes nearly 1.5 seconds. While reposing, SNARF performs this operation for every given pose, whereas INS requires mesh extraction only once. *Reposing the extracted mesh INS takes 0.13 seconds for an inference pass, which is an order of magnitude faster than SNARF.*

4.4. Ablations

We perform numerous ablations of our INS setup, the results of which are summarized in Table 3.

Multiplying Pose and Space Embeddings is important. Reformulating the pose conditioning by simple concatenation, i.e. $[\mathbf{e}_z, \mathbf{e}_\theta]$ instead of multiply-then-concatenate, i.e. $[\mathbf{e}_z, \mathbf{e}_z \odot \mathbf{e}_\theta]$ leads to a significant performance drop of $\sim 11\%$ points in both metrics (Row 2).

\mathbf{H}_c contributes much more than \mathbf{H}_d . The invertible networks do not contribute equally to the performance. Removing the canonical space PIN \mathbf{H}_c leads to a sharp drop in IoU Surface by 4.94%, when compared to removing the deformed space INN \mathbf{H}_d which drops performance by only 0.17% points (Rows 5,6). We attribute this partly to the fact that editing an LBS-deformed mesh introduces the additional complexity of resolving part-wise correspondences, such as locating the new positions of joints and limbs.

Replacing SIREN positional embedding with MLP hurts. IoU Surface drops by 3.16% if the learned sinusoidal embeddings are replaced by simple MLP layers. This happens as fine surface details such as cloth wrinkles get blurred when using MLP, which is prevented by SIREN.

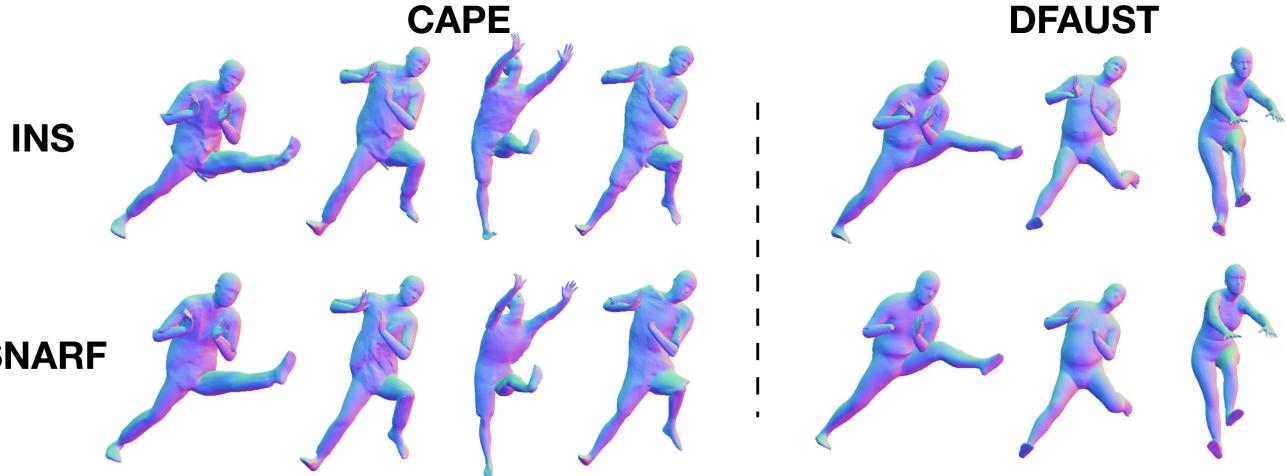


Figure 7. Comparison of INS and SNARF models undergoing extreme pose animation from PosePrior dataset.

PINs without rotation perform slightly worse. Removing 2D operations from our PINs leads to a drop in IoU Surface by 0.92%. This happens because twisting deformations in the surface have to be represented by only displacements, which has previously been shown to be difficult to learn [42].

Simply using PINs without LBS performs worse. Entirely removing the differential LBS module and relying solely on PINs to capture the full articulate motion results in a huge drop of 32% on both metrics (Row 7).

4.5. Qualitative Analysis

INS can represent finer details compared to SNARF. In Figure 6, we demonstrate a subject in a challenging novel pose from the CAPE dataset. We find that SNARF (left-most) is unable to capture fine details of cloth wrinkles, while also missing fingers as highlighted by the markers. Whereas SNARF-NC (second left) struggles with LBS artefacts such as volume loss by shrinking the arched back (highlighted), and displaying candy-wrapper effects. Finally, INS (second right) is able to capture much sharper local details around the body joints, such as around the waist and neck.

PINs can represent pose-varying deformations well. In Figure 8, we tease apart the edits made by solely PIN H_c in the canonical space (displayed in the top row) given two unseen target poses (shown in the bottom row). As highlighted in the figure, we find that PIN learns to introduce pose-varying deformations such as raising cloth outlines around the neck and shoulder joint, introducing dress wrinkles at near extremities, and even adjusting limbs such as orienting feet.

Extreme poses from PosePrior. The most loose-fitted sequence in CAPE is of subject 00375 wearing a blazer

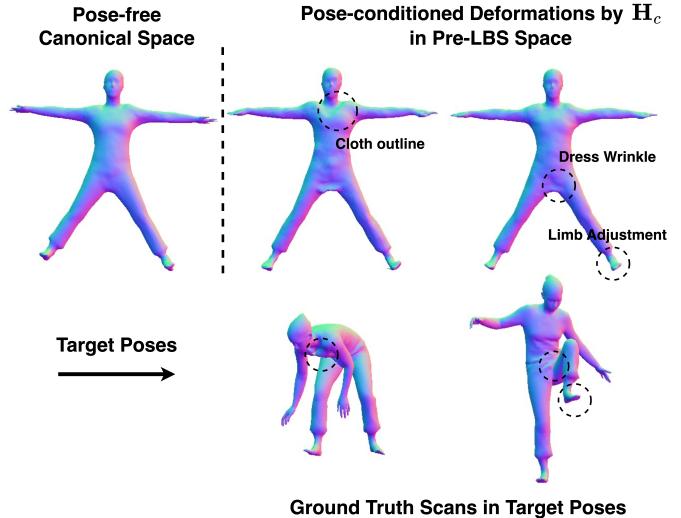


Figure 8. Pre-LBS space deformations. We show the edits made by PIN H_c before LBS operation for two novel target poses. Best viewed under zoom.

and trouser, we show it animated in extreme poses from Pose Prior in first two columns of Figure 7. We also visualize other subjects (both clothed and naked) in extreme poses as well. We find that INS produces much more realistic cloth deformations compared to SNARF.

5. Conclusion

In this work, we presented an invertible, end-to-end differentiable, and trainable pipeline called Invertible Neural Skinning for reposing humans. For this, we built a Pose-conditioned Invertible Network (PIN) that can handle non-linear surface deformations of clothes and skin well, while also retaining correspondences across different poses. By

placing two PINs around a differentiable LBS network and using a pose-free occupancy network, we created INS. We show that INS outperforms previous methods on clothed humans, while staying competitive on simpler and minimally clothed humans. Since reposing with our method requires the expensive mesh extraction exactly once, INS provides a speed-up of an order of magnitude compared to previous methods when animating long pose sequences.

Future Work. While training INS, the correspondence search performed by the differentiable LBS module often becomes a bottleneck, and future works can explore the possibility of eliminating this module completely — by learning its behavior from data. Furthermore, the occupancy network can be replaced with a neural representation that can handle texture and lightning and thus learn directly from 2D images and videos instead of raw scans.

A. Implementation Details

A.1. Sampling Points

Following SNARF, we sample 200K points at every frame of the sequence. Half of these points (100K) are near the mesh (scan) surface, which are obtained by first sampling points on the mesh surface via Poisson disk sampling and followed by displacement with isotropic Gaussian noise (of $\sigma = 0.01$). Remaining half (100K) points are sampled uniformly within a bounding box scaled to 110% of the original bounding box.

A.2. Hyperparameters and Training Details

We trained all our models on a single Tesla V100 GPU for 250 epochs, which took nearly 40 hours on average. We used a learning rate of $1e-4$ to train the PINs, while using a learning rate of $1e-3$ for remaining modules. We used Adam [25] optimizer, with a linear warmup and no learning rate decay. PyTorch [44] is used for all the experiments. Please refer to Table 4 for full list.

A.3. Metrics

Given set of sampled points \mathbf{P} to be evaluated, we can represent the joint tuple of any point, its ground truth occupancy (which can be either 0 or 1), and predicted occupancy as $(\mathbf{p}_d^i, g^i, h^i) \forall \mathbf{p}_d^i \in \mathbf{P}$ respectively. Then Intersection over Union (IoU) can be computed as follows:

$$\text{IoU} = \sum_{\mathbf{p}_d^i \in \mathbf{P}} \frac{g^i \cap h^i}{g^i \cup h^i} \quad (20)$$

To convert predicted probability to binary occupancy, we simply check if it is greater than 0.5. IoU Bounding Box operates with points sampled uniformly in the space, whereas Surface IoU operates with points sampled close to the body as described in Section A.1.

B. Data

CAPE. CAPE originally contains 15 subjects, with each subject wearing 1-6 different types of clothing, and performing 3-74 different actions. On average, it contains nearly 249 frames for every *subject-cloth* pair. Due to high variance as well as high number of *subject-cloth* pairs, we use a subset of CAPE which contains 15 sequences of 13 subjects containing all 8 different types of clothings. A clothing in CAPE is denoted by a joint string of upper and lower body garment, for example, a subject wearing a blazer and pants is annotated as *blazerlong*, and so on.

C. Invertible Neural Network

C.1. Initialization

We found that initializing the Pose-conditioned Invertible Networks (PINs) as identity modules stabilizes training, and allows the LBS network to train better. For this, we initialize the weights and biases of the last layer of the operation maps \mathbf{m}_r and \mathbf{m}_t (shown in Figure 2) as zeros.

C.2. Volume Preservation

Previous works in INNs [15, 16, 26] operating on high-dimensional (≥ 512 -d) spaces constrained the Jacobian between input and output to an triangular matrix. This ensured that the Jacobian determinant, used for density modeling, was not expensive to compute. Determinant of a triangular matrix is simply multiplication of its diagonal. However, this prevented these works from using 2D operators such as rotation. We note that such a requirement is unnecessary in our setting, where Jacobian determinant is not needed. Additionally, using rotations also helps to preserve volume between the input and output spaces.

Next, we show that our PINs consisting of only rotations and translations are volume preserving. Note that to show a transform is volume preserving it is sufficient to show that the determinant of the Jacobian of this transform is one. From Equations 5 and 10, we can express the transform represented by a single 2D coupling layer as:

$$\begin{aligned} x' &= x \cos(\gamma_{xy}) - y \sin(\gamma_{xy}) + t_x \\ y' &= x \sin(\gamma_{xy}) + y \cos(\gamma_{xy}) + t_y \\ z' &= z \end{aligned}$$

Then Jacobian of this transform becomes:

$$\mathbf{J} = \begin{bmatrix} \cos(\gamma_{xy}) & -\sin(\gamma_{xy}) & 0 \\ \sin(\gamma_{xy}) & \cos(\gamma_{xy}) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (21)$$

And determinant of the above Jacobian is one, i.e. $|\mathbf{J}| = 1$. Since, our PINs are composed of chaining together such coupling layers described in Equation 13, the overall determinant is also one. Hence volume is preserved within PINs.

#	Hyperparameters	Value	#	Hyperparameters	Value
1	No. of Parameters in INS	1.80M	2	No. of Coupling Layers in $\mathbf{H}_d/\mathbf{H}_c$	18
3	No. of Parameters in PIN $\mathbf{H}_d/\mathbf{H}_c$	0.41M	4	No. of Parameters in Occupancy Network \mathbf{O}	0.46M
5	No. of Parameters in LBS Network	53K	6	No. of Parameters in Bone Encoder	0.46M
7	Pose Embedding Dimension	120	8	Space Embedding Dimension	120
9	Space and Pose Embedding	240	10	PIN input and output Dimension	3
11	Number of epochs	250	12	Optimizer	Adam
13	Batch size (DFAUST/CAPE)	12/8	14	Learning rate (INNs/Rest)	1e-4/1e-3
15	Warm-up learning rate factor	0.2	16	Warm-up iterations	2400
17	No. of points per batch	60000	18	Gradient clipping (L-2 Norm)	4.0

Table 4. Hyperparameters and Training configuration to train INS.

D. Gradients

Training INS requires calculating gradients of the Binary Cross Entropy (BCE) loss \mathcal{L}_{bce} (Equation 18), with respect to all the components. Let the weights of PINs \mathbf{H}_c , \mathbf{H}_d , the LBS network \mathbf{w}_{lbs} , and the Occupancy network \mathbf{O} be denoted with σ_c , σ_d , σ_{lbs} , and σ_o respectively. Backpropagating through the occupancy network \mathbf{O} and the PIN \mathbf{H}_c is straightforward:

$$\frac{\partial \mathcal{L}_{bce}}{\partial \sigma_o} = \frac{\partial \mathcal{L}_{bce}}{\partial o} \cdot \frac{\partial o}{\partial \mathbf{O}(\mathbf{p}_c)} \cdot \frac{\partial \mathbf{O}(\mathbf{p}_c)}{\partial \sigma_o} \quad (22)$$

$$\frac{\partial \mathcal{L}_{bce}}{\partial \sigma_c} = \frac{\partial \mathcal{L}_{bce}}{\partial \mathbf{O}(\mathbf{p}_c)} \cdot \frac{\partial \mathbf{O}(\mathbf{p}_c)}{\partial \mathbf{H}_c(\mathbf{q}_c^*)} \cdot \frac{\partial \mathbf{H}_c(\mathbf{q}_c^*)}{\partial \sigma_c} \quad (23)$$

where o is the predicted occupancy. While the gradients for LBS network \mathbf{w}_{lbs} and second PIN \mathbf{H}_d are:

$$\frac{\partial \mathcal{L}_{bce}}{\partial \sigma_{lbs}} = \frac{\partial \mathcal{L}_{bce}}{\partial \mathbf{H}_c(\mathbf{q}_c^*)} \cdot \frac{\partial \mathbf{H}_c(\mathbf{q}_c^*)}{\partial \mathbf{q}_c^*} \cdot \frac{\partial \mathbf{q}_c^*}{\partial \sigma_{lbs}} \quad (24)$$

$$\frac{\partial \mathcal{L}_{bce}}{\partial \sigma_d} = \frac{\partial \mathcal{L}_{bce}}{\partial \mathbf{q}_c^*} \cdot \frac{\partial \mathbf{q}_c^*}{\partial \mathbf{H}_d(\mathbf{p}_d^t)} \cdot \frac{\partial \mathbf{H}_d(\mathbf{p}_d^t)}{\partial \sigma_d} \quad (25)$$

where \mathbf{q}_c^* is the root of the Equation 17, and \mathbf{p}_d^t is the input point. Pytorch’s automatic differentiation can handle the gradients in Equations 22 and 23. However, to obtain gradients w.r.t. \mathbf{q}_c^* implicit differentiation is required, similar to SNARF:

$$\begin{aligned} & \text{lbs}(\mathbf{q}_c^*, \theta^t) - \mathbf{p}_d^t = \mathbf{0} \\ \Leftrightarrow & \frac{\partial \text{lbs}(\mathbf{q}_c^*, \theta^t)}{\partial \sigma_{lbs}} + \frac{\partial \text{lbs}(\mathbf{q}_c^*, \theta^t)}{\partial \mathbf{q}_c^*} \cdot \frac{\partial \mathbf{q}_c^*}{\partial \sigma_{lbs}} = \mathbf{0} \\ \Leftrightarrow & \frac{\partial \mathbf{q}_c^*}{\partial \sigma_{lbs}} = - \left(\frac{\partial \text{lbs}(\mathbf{q}_c^*, \theta^t)}{\partial \mathbf{q}_c^*} \right)^{-1} \cdot \frac{\partial \text{lbs}(\mathbf{q}_c^*, \theta^t)}{\partial \sigma_{lbs}} \end{aligned} \quad (26)$$

And we can find gradients of \mathbf{q}_c^* with respect to \mathbf{q}_d^t as follows:

$$\begin{aligned} & \text{lbs}(\mathbf{q}_c^*, \theta^t) - \mathbf{p}_d^t = \mathbf{0} \\ \Leftrightarrow & \frac{\partial \text{lbs}(\mathbf{q}_c^*, \theta^t)}{\partial \mathbf{q}_c^*} \cdot \frac{\partial \mathbf{q}_c^*}{\partial \mathbf{H}_d(\mathbf{p}_d^t)} + \mathbf{1} = \mathbf{0} \\ \Leftrightarrow & \frac{\partial \mathbf{q}_c^*}{\partial \mathbf{H}_d(\mathbf{p}_d^t)} = - \left(\frac{\partial \text{lbs}(\mathbf{q}_c^*, \theta^t)}{\partial \mathbf{q}_c^*} \right)^{-1} \end{aligned} \quad (27)$$

E. Miscellaneous Failed Experiments

In order to help with a future research in this direction we list several ideas that have been tried in our project, which however did not improve performance.

E.1. Invertible Residual Layers

Idea. Beyond the invertible space-splitting layers, we also experimented with using invertible residual layers [18, 24]. These layers operate by limiting the Lipschitz constant of the residual branch, which has to be less than one in order to guarantee invertibility. Inversion of these layers can be achieved using a fixed-point iteration method, with convergence rate exponential in the number of iterations.

Outcome. We tried chaining residual layers with coupling layers alternately, and also placing them in the start and end of the invertible networks. However, these setups performed close or worse than without using residual layers, and were slower due to the expensive inversion pass.

E.2. Coupling Layers with Scales

Idea. Previous works CaDeX [28], and NeuralParts [43] utilized invertible networks with scale and translation operations, following the architecture proposed in RealNVP [16]. Such a transform can be represented as:

$$\begin{aligned} x' &= x \exp(s_x) + t_x \\ y' &= y \exp(s_y) + t_y \\ z' &= z \end{aligned}$$

Outcome. When using these layers in our experiments we encountered the following difficulties:

- **Unstable Training.** Floating point overflows occurred frequently during training due to the exponential scaling term. Even after carefully tuned gradient clipping and learning rate schedules, we encountered frequent experiment failures.
- **Squashing Effect.** Since the scaling operator can lead to very large outputs from INN, generally a sigmoid squashing layer is used at the end to restrict the input to a fixed range that matches output distribution. However, due to this sigmoid layer, the INN can no longer be initialized as an Identity layer, even when all the rotations are identity and translations are zero. This leads to a squashing artefacts in outputs.
- **Non-volume Preserving.** The Jacobian of these layers [28] with scaling is not one, previously derived Equation 21. Due to this additional regularization is needed for training.

E.3. Pose-conditioned 3D rotation and translation layers

Idea. We tried learning pose-conditioned global 3D rotation and translation layers. We implemented it similar to the coupling layers to predict rotation and translation parameters, but without any space conditioning.

Outcome. We did not find significant gains using this, and decided against using them in the final version as they had a big memory footprint. These layers often rotate the canonical space creating issues during mesh extraction.

F. Visuals ([video link](#))

We place all the qualitative results in this [video](#), and discuss its contents below.

[Video Part-1] Pose-varying deformations in INS. In the first part of the video, we visualize the deformations introduced by PINs \mathbf{H}_c and \mathbf{H}_d under varying target poses (shown in top right). Deformations introduced by \mathbf{H}_c are shown in the top-middle part, whereas those introduced by \mathbf{H}_d are shown in the bottom-left part shaded in green. We demonstrate that INS is able to handle complex deformations of clothing across poses.

[Video Part-2] Baseline Comparison. In the second part of the video, we compare our method INS against all the five baselines discussed in Section 4.2 of the main paper. While both the LBS baselines, and SNARF-NC suffers from artifacts, we see that INS performs much better than other methods.

[Video Part-3] INS Ablations. In the third part of the video, we visualize results from various ablations reported in Section 4.4 of the main paper. Here, we find that removing SIREN leads to an overly smooth surface, and removing the LBS network makes it harder for the network to learn limb movements correctly.

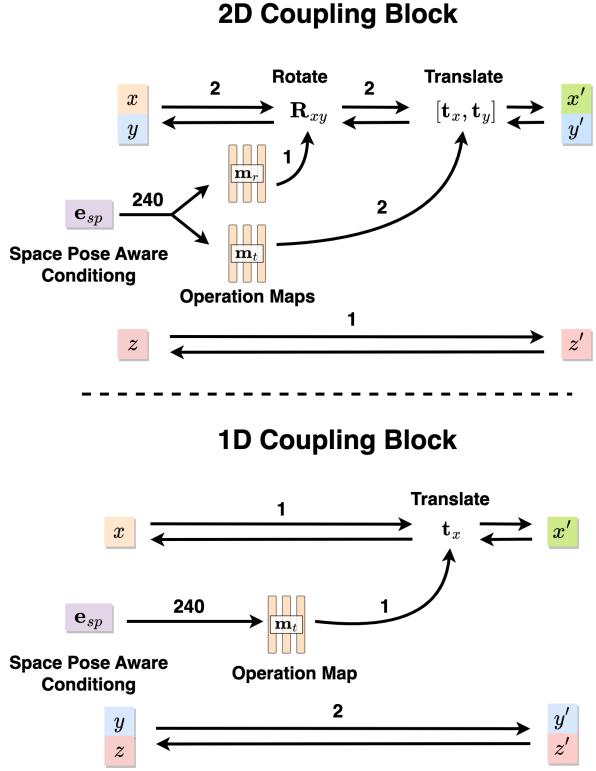


Figure 9. **1D and 2D Coupling Layers.** We show comparison between both types of layers used in PINs. The bidirectional arrows show invertible computations.

[Video Part-4] Texture Propagation. As INS can preserve correspondences across poses, it becomes possible to propagate mesh attributes such as texture across various time frames. We conducted an experiment to test this, where we applied texture to the pose-independent canonical mesh. Next, we propagated this texture through the INS network. We show the results of this experiment in the fourth (and last) part of the video. We found that the applied texture deformed realistically like clothing, while being consistent across all frames, and was free of jittering.

To contrast and compare with the above experiment, we conducted similar texture propagation using SNARF. Since, SNARF decodes a separate mesh at each time-step, we color this mesh using the same scheme for coloring INS canonical mesh above. Propagating this texture through the LBS block, we find that it frequently leads to jittery artefacts as the texture overflows across semantically different parts. For example, the texture patch E4 applied to the blazer in a particular frame, overflows onto pants in another frame, and so on.

F.1. 1D and 2D Coupling Layers

We visualize both 1D and 2D coupling layers together in Figure 9 for better understanding. In 1D case the space-pose aware conditioning gets conditioned on the 2D input,

#	Experiment	IoU Surf.	IoU BBox	Train Iter (sec)
1	SNARF-MLP	63.66	63.74	1.80
2	w/ Pose Mul.	<u>68.62</u>	<u>68.64</u>	1.80
3	SNARF-NC	66.10	66.11	1.16
4	INS	72.83	72.69	<u>1.34</u>

Table 5. Experiments with setup similar to IMAvatar [20].

which helps to improve expressiveness. In the 2D case, we can make edits on an entire 2D plane conditioned on the 1D input. We find out that these blocks provide complementary benefits, thus we utilize both of them in the final architecture.

G. Miscellaneous

G.1. Training speed of INS comparison.

Training the LBS module (Broyden’s method) takes nearly 0.70s, while both PIN modules combined take 0.09s per iteration. Accounting for loss computation and back-prop, single iteration of SNARF takes 1.16s, while INS takes 1.34s. Overall there is a 13.4% slowdown using INS.

H. Comparison with IMAvatar

We can model non-rigid deformations using MLP and leverage Broyden’s method in training, similar to IMAvatar [20].

Experiment: We conducted a preliminary study, where we used a second MLP to model pose-conditioned vertex offsets before the LBS module similar to IMAvatar. We reused the clothed sequence from ablation study, and name this experiment as **SNARF-MLP**. We also added two tricks from INS to make this setup work. First, we zero initialized the last layer of offset MLP; second, we used the space-pose aware conditioning described in Equation 19 (Section 3.2).

Quantative (Table 1): We found SNARF-MLP performs subpar compared to INS (Row 1 vs 4), while training much slowly due to multiple MLP runs (44% slowdown). Moreover, we find our pose-conditioning to boosts performance (Row 2), while not using zero init. leads to divergence.

Qualitative (Fig.1, RHS): Visually, we found SNARF-MLP to result in unnatural deformations (highlighted). We hypothesize, this could be due to the complex landscape of the resulting function which might make Broyden’s vulnerable to local minima. Thus, points close to each other, which ideally should have close solutions, converge to different local minima producing bending artefacts. In contrast, PINs optimized with SGD can avoid local minima [19].



Figure 10. Visuals showing non-rigid deformations modeled using a second MLP similar to IMAvatar [20].

Subject	IoU Surface			IoU Bounding Box		
	SNARF	SNARF-NC	INS (ours)	SNARF	SNARF-NC	INS (ours)
03375	<u>70.16%</u>	66.1%	71.13%	62.57%	<u>70.24%</u>	71.02%
50007	90.28%	83.9%	<u>86.63%</u>	97.77%	<u>96.16%</u>	96.11%
50022	<u>92.19%</u>	88.09%	92.58%	98.05%	96.68%	98.12%
50026	91.13%	80.54%	<u>89.26%</u>	97.67%	94.37%	<u>97.13%</u>
50004	89.6%	85.4%	<u>88.96%</u>	97.48%	96.3%	<u>97.21%</u>
50009	<u>87.05%</u>	83.87%	<u>85.77%</u>	95.89%	<u>94.63%</u>	93.47%
50021	<u>89.76%</u>	87.26%	90.46%	<u>96.79%</u>	95.58%	96.95%
50025	<u>90.95%</u>	86.12%	91.59%	<u>97.35%</u>	95.82%	97.55%
50027	89.54%	<u>86.9%</u>	85.91%	96.74%	<u>95.79%</u>	93.75%
50002	89.25%	84.41%	<u>85.98%</u>	97.55%	96.67%	<u>97.02%</u>
50020	90.31%	85.69%	<u>88.74%</u>	96.85%	95.15%	<u>96.23%</u>
Average	90.01%	85.22%	<u>88.59%</u>	97.21%	95.72%	<u>96.35%</u>

Table 6. Quantitative Results on Minimally Clothed Humans. On DFAUST, INS outperforms SNARF-NC by a large margin while performing competitively with SNARF, and being order magnitude faster at reposing.

I. Subjectwise result breakdown

Expanding on the results from main draft, we also report per-subject breakdown on DFAUST (Table 6) and CAPE (Table 7) datasets.

References

- [1] Thimo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imGHUM: Implicit generative models of 3d human shape and articulated pose. In *Int. Conf. Comput. Vis.*, 2021. 1
- [2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416, 2005. 2

Subject	Clothing	IoU Surface					IoU Bounding Box				
		AVG-LBS	FIRST-LBS	SNARF	SNARF-NC	INS (ours)	AVG-LBS	FIRST-LBS	SNARF	SNARF-NC	INS (ours)
50002	longshort	84.80%	87.89%	47.34%	96.56%	97.50%	63.86%	66.42%	85.41%	84.02%	89.57%
03375	blazerlong	62.63%	53.91%	70.16%	66.1%	72.83%	62.57%	54.05%	70.24%	66.11%	72.69%
00215	poloshort	64.91%	57.27%	73.1%	65.68%	72.22%	64.92%	57.55%	73.1%	65.8%	72.35%
00096	shirtlong	63.48%	52.07%	75.85%	67.92%	73.48%	63.55%	52.15%	75.77%	67.97%	73.56%
00096	shirtshort	59.05%	54.67%	75.32%	63.59%	74.89%	59.12%	54.7%	75.16%	63.61%	74.96%
00096	jerseyshort	62.0%	56.63%	73.28%	64.28%	74.61%	62.02%	56.45%	73.19%	63.98%	74.37%
00134	longlong	65.98%	61.5%	73.96%	67.65%	78.97%	65.91%	61.45%	73.96%	67.73%	78.93%
03223	shortshort	74.74%	60.66%	81.42%	77.56%	82.63%	74.81%	60.67%	81.33%	77.66%	82.76%
03331	longshort	70.3%	65.2%	77.13%	75.12%	77.12%	70.5%	65.52%	77.03%	74.5%	77.22%
00127	shortlong	73.84%	65.7%	72.48%	70.0%	73.2%	74.09%	65.87%	72.31%	69.9%	73.33%
02474	longshort	60.87%	54.37%	70.16%	62.39%	71.64%	60.94%	54.41%	70.14%	62.26%	71.76%
03284	longshort	64.68%	58.5%	67.04%	65.76%	68.79%	64.68%	58.21%	66.87%	65.44%	68.77%
00032	longshort	64.05%	59.63%	69.23%	64.47%	69.04%	64.51%	59.67%	69.21%	64.63%	69.14%
00122	shortlong	56.85%	45.74%	64.67%	60.98%	64.85%	57.01%	45.85%	64.56%	60.87%	65.06%
03394	longlong	62.56%	52.11%	69.43%	66.16%	71.55%	62.6%	52.25%	69.27%	66.1%	71.36%
00159	longshort	69.27%	63.22%	70.31%	65.64%	71.1%	69.52%	63.76%	70.35%	65.19%	71.58%
Average		65.01%	57.41%	72.24%	66.89%	73.13%	65.12%	57.5%	72.17%	66.78%	73.19%

Table 7. **Quantitative Results on Clothed Humans.** We find our approach INS outperforms all methods when averaged across 15 runs, on both IoU Surface and IoU Bounding Box metrics.

- [3] Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W. Pellegrini, Ralf S. Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks, 2018. [2](#)
- [4] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, 2020. [2](#)
- [5] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In *Asian Conference on Computer Vision*, pages 3–19. Springer, 2018. [2](#)
- [6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Eur. Conf. Comput. Vis.*, 2016. [1](#)
- [7] Charles G Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, pages 19(92):577–593, 1965. [4](#)
- [8] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10754–10764, October 2021. [2](#)
- [9] Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang. Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022. [2](#)
- [10] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J. Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars, 2022. [2](#)
- [11] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. *arXiv preprint arXiv:2104.03953*, 2021. [1, 2, 3, 4, 6](#)
- [12] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. [1](#)
- [13] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *The European Conference on Computer Vision (ECCV)*. Springer, August 2020. [1](#)
- [14] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 612–628. Springer, 2020. [1, 2, 3](#)
- [15] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. [1, 2, 4, 9](#)
- [16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. [1, 2, 4, 9, 10](#)
- [17] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksandr Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. [2](#)
- [18] Behrmann et al. Invertible residual networks. In *ICML*, 2019. [2, 10](#)
- [19] Keskar et al. On large-batch training for deep learning: Generalization gap and sharp minima, 2017. [12](#)
- [20] Zheng et al. I M Avatar: Implicit morphable head avatars from videos. In *CVPR*, 2022. [2, 12](#)
- [21] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution es-

- timation. In *International Conference on Machine Learning*, pages 881–889. PMLR, 2015. 2
- [22] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2
- [23] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: animatable reconstruction of clothed humans. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3090–3099. IEEE, 2020. 2
- [24] Jörn-Henrik Jacobsen, Arnold W.M. Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. In *International Conference on Learning Representations*, 2018. 2, 10
- [25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015. 9
- [26] Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018. 1, 2, 4, 9
- [27] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2
- [28] Jiahui Lei and Kostas Daniilidis. Cadex: Learning canonical deformation coordinate space for dynamic surface representation via neural homeomorphism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 10, 11
- [29] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors, 2022. 2
- [30] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2
- [31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 2, 3, 6
- [32] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987, Anaheim, California, USA, July 27-31, 1987*, pages 163–169. ACM, 1987. 2
- [33] Qianli Ma, Jinlong Yang, Michael J. Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *2022 International Conference on 3D Vision (3DV)*, September 2022. 2
- [34] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3d people in generative clothing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6468–6477. IEEE, 2020. 6
- [35] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Int. Conf. Comput. Vis.*, 2019. 6
- [36] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4460–4470. Computer Vision Foundation / IEEE, 2019. 1, 2, 3, 7
- [37] Marko Mihajlovic, Yan Zhang, Michael J. Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 1, 2
- [38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 2
- [39] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. DeepHandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pages 440–455. Springer, 2020. 2
- [40] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: sparse trained articulated human body regressor. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 598–613. Springer, 2020. 2
- [41] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *arXiv preprint arXiv:1705.07057*, 2017. 2
- [42] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields, 2020. 8
- [43] Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, and Sanja Fidler. Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3204–3215, 2021. 2, 10
- [44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 9
- [45] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and

- Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [46] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. *arXiv preprint arXiv:2105.02872*, 2021. 2
- [47] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2
- [48] Chengping Rao, Hao Sun, and Yang Liu. Physics informed deep learning for computational elastodynamics without labeled data, 2020. 2
- [49] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):1–17, 2017. 2
- [50] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morigima, Hao Li, and Angjoo Kanazawa. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2304–2314. IEEE, 2019. 1
- [51] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 81–90. IEEE, 2020. 1
- [52] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 1, 2
- [53] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yanpei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *CVPR*, 2022. 2
- [54] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 4
- [55] Shih-Yang Su, Timur M. Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. *ArXiv*, abs/2205.01666, 2022. 1
- [56] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. *arXiv preprint arXiv:2106.11944*, 2021. 1, 2
- [57] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdf. In *European Conference on Computer Vision*, 2022. 1, 2
- [58] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humanerf: Free-viewpoint rendering of moving people from monocular video, 2022. 2
- [59] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022. 2
- [60] Hongyi Xu, Thimo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion, 2021. 2
- [61] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: generative 3D human shape and articulated pose models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6183–6192. IEEE, 2020. 2
- [62] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022. 2
- [63] Hao Zhao, Jinsong Zhang, Yu-Kun Lai, Zerong Zheng, Yingdi Xie, Yebin Liu, and Kun Li. High-fidelity human avatars from a single rgb camera. In *CVPR*, 2022. 2
- [64] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling, 2022. 2
- [65] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J. Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2