Neural-GIF: Neural Generalized Implicit Functions
for Animating People in Clothing

Garvita Tiwari [1,2]   Nikolaos Sarafianos [3]   Tony Tung [3]   Gerard Pons-Moll[1,2]

[1]University of Tübingen, Germany,   [2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

[3]Facebook Reality Labs, Sausalito, USA

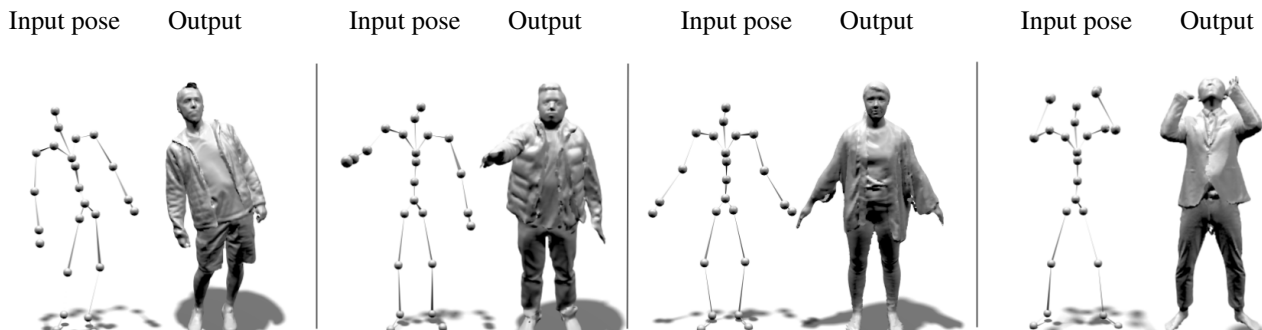`gtiwari@mpi-inf.mpg.de`, {`nsarafianos`, `tony.tung`}`@fb.com`, `gerard.pons-moll@uni-tuebingen.de`

Figure 1. We present *Neural Generalized Implicit Functions* (Neural-GIF), to animate people in clothing as a function of body pose. Neural-GIF learns directly from scans, models complex clothing and produces pose-dependent details for realistic animation. We show for four different characters the query input pose on the left (illustrated with a skeleton) and our output animation on the right.

Abstract

We present Neural Generalized Implicit Functions *(Neural-GIF), to animate people in clothing as a function of the body pose. Given a sequence of scans of a subject in various poses, we learn to animate the character for new poses. Existing methods have relied on template-based representations of the human body (or clothing). However such models usually have fixed and limited resolutions, require difficult data pre-processing steps and cannot be used with complex clothing. We draw inspiration from template-based methods, which factorize motion into articulation and non-rigid deformation, but generalize this concept for implicit shape learning to obtain a more flexible model. We learn to map every point in the space to a canonical space, where a learned deformation field is applied to model non-rigid effects, before evaluating the signed distance field. Our formulation allows the learning of complex and non-rigid deformations of clothing and soft tissue, without computing a template registration as it is common with current approaches. Neural-GIF can be trained on raw 3D scans and reconstructs detailed complex surface geometry and deformations. Moreover, the model can generalize to new poses.*

We evaluate our method on a variety of characters from different public datasets in diverse clothing styles and show significant improvements over baseline methods, quantitatively and qualitatively. We also extend our model to multiple shape setting. To stimulate further research, we will make the model, code and data publicly available at [1].

1. Introduction

Human avatars enable numerous applications related to augmented and virtual reality, such as telepresence for enhanced communication and entertainment, and have been instrumental to reconstruct and perceive people in images [2, 3, 11, 22, 62, 61]. Human shape deforms according to articulation, soft-tissue and non-rigid clothing dynamics, which make realistic animations extremely challenging.

State-of-the-art body models [29, 43, 61] typically learn to deform a *fixed topology template*, usually using linear blend skinning to model articulation, and blendshapes to model non-rigid effects [29], even including soft-tissue [45] and clothing [31, 42, 52]. The use of a fixed template limits the type of clothing and dynamics that can be modeled. For example, it would be difficult to model the subjects in

Fig. 1 with one or more predefined templates. Furthermore, every type of deformation (soft-tissue or clothing) requires a different model formulation. Additionally for training a model, 3D/4D scans need to be brought into correspondence [29, 45], which is a challenging task especially for clothing [44], and not even well defined when garments vary in topology [10]. Recent works have leveraged implicit function representation to reconstruct human shape from images [49, 50] or 3D point-clouds [8, 15]. These reconstructions are however static and not animatable.

In this work, we propose a new model, called Neural Generalized Implicit Functions (Neural-GIF) to animate people in clothing as a function of body pose. We demonstrate that we can model complex clothing and body deformations at a quality not easily achieved with a traditional fixed template-based mesh representations. In contrast to most prior work, we learn pose-dependent deformations without the need of registration of any pre-defined template, which degrades the resolution of the observations, and is a notoriously complex step prone to inaccuracies for complex clothing [44]. Instead we solely require the pose of the input scans as well as the SMPL shape parameter (β). Another key advantage of our method is that it can represent different topologies using the exact same formulation –we show how to animate jackets, coats, skirts and soft-tissue of undressed humans. Neural-GIF consists of a neural network to approximate the signed distance field (SDF) of the posed surface. Naively learning to predict SDF from pose is hard. Instead, we draw inspiration from template-based methods which factorize motion into articulation and non-rigid deformation, but generalize this concept for implicit shape learning. Specifically, we learn to map every point surrounding the surface to a canonical space, where a learned deformation field is applied to model non-rigid effects, before evaluating the SDF. Our model (and its name) is inspired by the seminal paper [53], which shows that a wide variety of shapes can be obtained by simply applying deformation fields to a base implicit shape. The advantage of Neural-GIF is that the network can more easily learn a base shape in canonical space, which can be deformed. In summary, our contributions are:

- Neural-GIF, an implicit based re-posable character, which can be directly learned from 3D scans. Our model can represent complex character/clothing scans of varied topology and geometry.

- We introduce a canonical mapping network, which learns continuous skinning field in 3D space and unpose 3D points surrounding the scan to a canonical T-pose, without explicit supervision.

- We introduce a *displacement* field network, which shifts points in the canonical space before evaluating the SDF, yielding in fine details and deformation.

We test our method on a variety of scans originating from different datasets and provide extensive quantitative and qualitative comparisons. We also extend our formulation to multiple shape setting, by adding a shape dependent displacement field network. Experiments demonstrate that our method generalizes to new poses, model complex clothing, and is significantly more robust and more detailed than existing methods [51, 36, 18].

2. Related Work

Human and Clothing models: The problem of animating clothing/characters has been tackled using Physics-Based Simulation (PBS) methods for a long time. 3D meshes are animated using PBS governed by some fundamental forces and collision detection. PBS methods require designing a garment/mesh template and manual intervention for finding suitable physical parameters. Past work has explored ways to automate this limitation by trying to infer these physical parameters mechanically from data [35, 59], or use videos to infer them [48, 54]. PBS methods are also computationally expensive and the complexity increases for realistic animations. Several approaches generate realistic and high-frequency details by computing a high resolution mesh from a base coarse mesh using constraint-based optimization [38, 47]. One way of animating characters is by learning a transformation between pose-shape and the body surface and one can generate characters in new pose and body shape using this linear transformation [5, 29, 39]. Since such body models do not model personalized and clothing details, various methods [4, 10, 44, 26] have explored ways to extend the underlying body model [29], by predicting per-vertex displacements for personalized and clothing features. However, they do not take pose-dependent deformations into account. Dyna [45], extends the body model to predict soft-tissue dynamics caused by motion by learning a second-order auto-regressive model. With the advances in deep learning and 3D data acquisition [56, 57, 33], data-driven models of clothing and humans [62] are becoming increasingly popular. One key aspect in modeling human and clothing is articulation and pose-dependent non-rigid deformations. A significant amount of prior work models these deformations conditioned on the body shape-pose [6, 20, 52], the body pose [27], the body shape-clothing size/style [55, 60] and the body shape-pose-clothing style [42]. All these methods use a template-specific model, *i.e.* either they are learned from offline PBS data [6, 20, 42, 52, 60, 21, 58] or learned from template-specific registration of scans [27, 31, 55]. [65] generates fine details for any clothing styles, but requires low resolution meshes as input. This strictly limits the usability of such approaches to new clothing styles, especially to those with complex geometries and does not follow the body topology. There is no existing work that can model
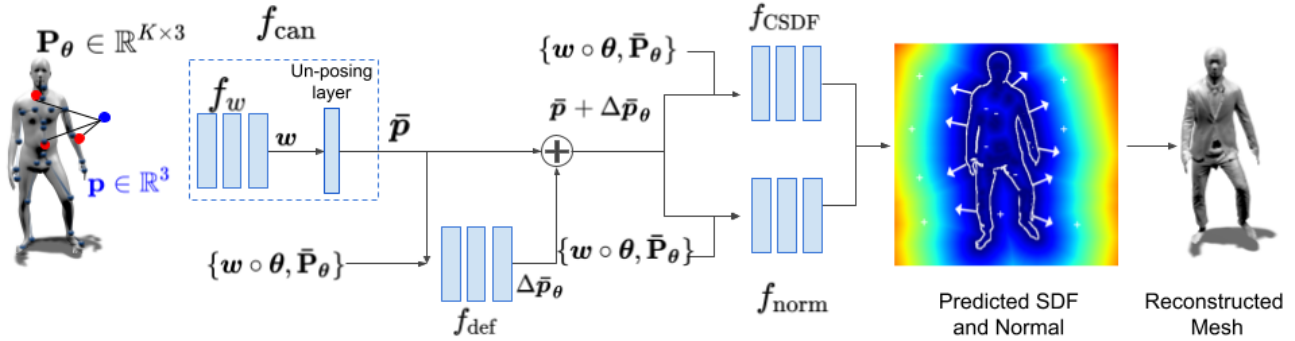
Figure 2. We propose *Neural Generalized Implicit Function* (Neural-GIF), which first maps every point p in 3D posed space to the canonical space by learning the associations between the point and the human body with f_w based on a pose encoding ($\mathbf{P}_{\theta}$) as input. Then we predict a displacement field model (f_{def}) learns a pose-dependent deformation field in canonical space ($\Delta\bar{p}_{\theta}$), conditioned on weighted pose parameters ($w \circ \theta$) and pose encoding ($\bar{\mathbf{P}}_{\theta}$) in the canonical space. We add the displacement to the unposed point before evaluating the SDF using our canonical SDF network (f_{CSDF}). We also employ a normal prediction network (f_{norm}) to obtain results with fine-level details. Both f_{CSDF} and f_{norm} take weighted pose and pose encoding as input.

complex clothing like jackets, directly from scans.

Modeling using Implicit representations: There is a plethora of recent methods [15, 16, 34, 40] that explore neural implicit functions to represent 3D objects, including detailed articulated objects like human bodies and garments [15, 16] but do not model pose-conditioned occupancy. Corona *et al.* [17] introduces a generative model for garments with different topology, but doesn't model fine pose dependent deformations. Like us, NASA [18] extends the implicit surface formulation to represent articulated objects using a composition of part-based neural implicit functions. However the part-based representation in [18] yields artifacts at the interfaces and does not capture fine details. Similar to our work, concurrent works like LEAP [36], NiLBS [25] and SCANimate [51], propose a neural implicit model which extends the idea of skinning weights to volumetric skinning weights. Previous work like ARCH [23] and LoopReg [9] also extend skinning weights to 3D space, but they are not learned and instead are pre-determined based on nearest neighbour. Prior work like [28, 7] learn skinning weights for different topolgies using graph convolutions. Yang *et al.* [63], propose to predict shape, skeleton and skinning jointly from image or depth data to animate characters. Also, in [23], clothed humans are modeled as pixel-aligned implicit functions in the unposed canonical space but do not model pose-dependent effects. Božič *et al.* [14] model non rigid deformations using an implicit deformable shape via an identity specific deformation graph but their focus is tracking and reconstruction of non-rigid 3D objects. In [64], an implicit morphable model for faces is proposed, where shape is decoupled into an implicit reference shape and a deformation of this reference shape.

Another line of work in implicit scene representation comes from NeRF [37], where the focus is photorealistic rendering of a scene. Dynamic NeRF methods [41, 46] add the dynamic aspect to neural radiance fields, by first encoding the scene into canonical space, using a deformation field and then predicting density and view-dependent color values. Hence they can reconstruct and render novel images of objects under rigid and non-rigid motions from a single camera moving around the scene. In contrast, we learn a model which generalizes to new poses, learns from 3D scans, and integrates the SMPL kinematic skeleton.

In the seminal work of [53], a generalized formulation of implicit functions is introduced to obtain new shapes by deforming and transforming an implicit base shape, by deforming and displacing points. Instead of using predefined deformation matrices and displacement maps, we impose articulated structure from the SMPL body model [29], learn to map points to a canonical space, and learn a pose-dependent displacement field with neural networks.

3. Method

Our method, called Neural-GIF, is a neural model based on generalized implicit functions [53] for animating people in clothing. Given a pose(θ) as input, Neural-GIF predicts SDF, whose zero level set represents the character's surface $\mathcal{S}$. Without loss of generality, we use SMPL [29] for our pose representation. To train Neural-GIF we require sequences of 3D scans of a subject in a fixed clothing, and the corresponding SMPL parameters. We do not require non-rigid registration and un-posing as a pre-processing step [42, 44, 45, 27], which is a tedious task and prone to error. Since learning the mapping $\theta \mapsto \text{SDF}(\mathcal{S})$ is hard (large variations due to articulation, fine non-rigid deformations due to soft-tissue and clothing) and often results in missing limbs, we factorize the motion to learn a *deformable SDF in*

a canonical (unposed) space of the character. Three neural functions are composed to obtain the pose dependent SDF:

- **Canonical mapping network** (Sec. 3.1): maps every point in the posed 3D space to the canonical unposed space by learning point to human body associations.

- **Displacement field network** (Sec. 3.2): models non-rigid pose-dependent deformations (soft-tissue, cloth dynamics). Specifically, it predicts a continuous displacement field for points in the canonical space.

- **Canonical SDF** (Sec. 3.3): By composition, *Canonical SDF* network, takes the transformed point with the above networks, and a pose encoding as input to predict the desired *signed distance* for every query point. We additionally predict a surface normal field as a function of pose in canonical space to add realism in the results.

3.1. Canonical Mapping Network

Our canonical space is characterized by a T-pose of character. Learning in this space allows the network to discover a mean shape that can be deformed and articulated, which helps in pose generalization. Formally, the main task of the network is to map every point in the posed space (including off-surface points) to the canonical space $\boldsymbol{p} \mapsto \bar{\boldsymbol{p}}$. This requires associating the point $\boldsymbol{p}$ with the body, and inverting the articulated motion. While we could associate every point to its closest SMPL body model point, this is often ambiguous in occluded regions near the armpits or between the legs(See fig. 8). Consequently, we propose to learn this *association as a function of pose*.

Instead of simply using pose $\boldsymbol{\theta}$ as input, we use a body-aware pose representation. First, we take K joints the SMPL mesh ($\boldsymbol{j_i}, i \in \{1 \ldots K\}$), and then define a pose encoding $\mathbf{P_\theta} \in \mathbb{R}^{K \times 3}$, where $\mathbf{P}_{\boldsymbol{\theta},i} = ||\boldsymbol{p} - \boldsymbol{j_i}||$.

Given a query point in space and its pose encoding, the *canonical mapping network* predicts the blend weights $\boldsymbol{w} \in \mathbb{R}^K$ of query point $\boldsymbol{p}$ and transforms it to the canonical space using a differentiable un-posing layer:

$$\boldsymbol{w} = f_{\boldsymbol{w}}(\boldsymbol{p}, \mathbf{P_\theta}) \qquad (1)$$

$$\bar{\boldsymbol{p}} = \left(\sum_{i=1}^{K} \boldsymbol{w}_i \mathbf{B}_i\right)^{-1} \boldsymbol{p}, \qquad (2)$$

where $\mathbf{B}_i$ is the transformation matrix for joints $i \in \{1 \ldots K\}$, and $\boldsymbol{w}_i$ its blend-weight. Eq. (1) and (2) define the canonical mapping network $f_{\text{can}} : (\boldsymbol{p}, \mathbf{P_\theta}) \mapsto \bar{\boldsymbol{p}}$.

3.2. Displacement Field Network

A surface expressed implicitly as the zero level set $\{\boldsymbol{x} \in \mathbb{R}^3 \,|\, SDF(\boldsymbol{x}) = 0\}$ can be deformed to obtain a new surface by applying transformations to the points themselves. By applying a displacement $\boldsymbol{x}' = \boldsymbol{x} + \Delta\boldsymbol{x}$ we will effectively displace the zero level

set $\{\boldsymbol{x} \in \mathbb{R}^3 \,|\, SDF(\boldsymbol{x}') = 0\}$, and consequently the surface defined. To model subtle non-rigid deformations, we propose to explicitly learn a pose-dependent displacement field in the canonical space:

$$\Delta\bar{\boldsymbol{p}}_{\boldsymbol{\theta}} = f_{\text{def}}(\bar{\boldsymbol{p}}, \boldsymbol{w} \circ \boldsymbol{\theta}, \bar{\mathbf{P}}_{\boldsymbol{\theta}}) \qquad (3)$$

where $\boldsymbol{w} \circ \boldsymbol{\theta}$ is weighted SMPL pose parameter and $\bar{\mathbf{P}}_{\boldsymbol{\theta}}$ is pose encoding computed in the canonical space. We denote the space of pose encoding as $(\boldsymbol{w} \circ \boldsymbol{\theta}, \bar{\mathbf{P}}_{\boldsymbol{\theta}}) \in \mathcal{P}$. Note that these are *not per-vertex displacements*, but a *learned continuous displacement field* $\mathbb{R}^3 \mapsto \mathbb{R}^3$, and hence we are not restricted to a fixed mesh with fixed number of vertices.

3.3. Canonical SDF and Normal Prediction

Given points in canonical space and the pose encoding $(\boldsymbol{w} \circ \boldsymbol{\theta}, \bar{\mathbf{P}}_{\boldsymbol{\theta}}) \in \mathcal{P}$, we learn a signed distance field in canonical space. The canonical SDF ($f_{\text{CSDF}} : \mathbb{R}^3 \times \mathcal{P} \to \mathbb{R}$) is parameterized with a fully-connected neural network. As seen in Fig. 2, the network predicts SDF values for points in the canonical space with the following equation:

$$d^* = f_{\text{CSDF}}(\bar{\boldsymbol{p}} + \Delta\bar{\boldsymbol{p}}_{\boldsymbol{\theta}}, \boldsymbol{w} \circ \boldsymbol{\theta}, \bar{\mathbf{P}}_{\boldsymbol{\theta}}) \qquad (4)$$

where the canonical point $\bar{\boldsymbol{p}} = f_{\text{can}}(\boldsymbol{p}, \mathbf{P}_{\boldsymbol{\theta}})$, and the displacement $\Delta\bar{\boldsymbol{p}} = f_{\text{def}}(\bar{\boldsymbol{p}}, \boldsymbol{w} \circ \boldsymbol{\theta}, \bar{\mathbf{P}}_{\boldsymbol{\theta}})$ are computed with the canonical mapping and displacement field networks respectively. This formulation has several nice properties. First, the network can implicitly learn a canonical shape of clothing/human, without ever explicitly un-posing the scans. Note that if the motion consists exclusively of only articulation and displacement, then we could directly learn $f_{\text{CSDF}}(\bar{\boldsymbol{p}} + \Delta\bar{\boldsymbol{p}})$ without further conditioning (in practice we condition on pose to give the network the flexibility to learn shapes which can not be modeled via displacements from a canonical shape). Second, we can supervise Eq. (4) with distances computed in the original posed scans.

In addition to signed distances, we predict the normals for points near the surface, to generate more realistic looking results. We predict normals in canonical space, using f_{norm} and then transform it to pose space using blend weights, computed from Eq. (1) and $\mathbf{B}_i$:

$$\boldsymbol{n}^* = \sum_{i=1}^{K} \boldsymbol{w}_i \mathbf{B}_i f_{\text{norm}}(\bar{\boldsymbol{p}} + \Delta\bar{\boldsymbol{p}}_{\boldsymbol{\theta}}, \boldsymbol{w} \circ \boldsymbol{\theta}, \bar{\mathbf{P}}_{\boldsymbol{\theta}}) \qquad (5)$$

3.4. Multi-subject model using Neural-GIF

We extend our proposed framework for multi-subject setting, *i.e.* we learn multiple shapes in one single model. For this we take motivation from SMPL [29] model, where per-vertex shape dependent displacement is added and extend the idea to a continuous shape dependent displacement field. To incorporate shape dependent displacement field, we introduce another displacement network($f_{\text{def}-\boldsymbol{\beta}}$), which predicts continuous displacement field in $\mathbb{R}^3$, given shape

parameter($\boldsymbol{\beta}$) as input:

$$\Delta\bar{\boldsymbol{p}}_{\boldsymbol{\beta}} = f_{\text{def}-\boldsymbol{\beta}}(\bar{\boldsymbol{p}}, \boldsymbol{\beta}) \qquad (6)$$

We further modify our model by taking SMPL shape(β) as additional input to each sub-network and adding $\Delta\bar{\boldsymbol{p}}_{\boldsymbol{\beta}}$ to $\bar{\boldsymbol{p}} + \Delta\bar{\boldsymbol{p}}_{\boldsymbol{\theta}}$ before applying $f_{\text{CSDF}}(*)$ and $f_{\text{norm}}(*)$.

3.5. Training and Inference

Training: Each of the sub-networks is implemented using an MLP. Following [37], we use a positional encoding for query point $\boldsymbol{p}$. We follow a multi-stage training regime for faster convergence and stable training. First we pre-train the canonical mapping network (f_{can}) using SMPL meshes generated from the training motion sequences and initially supervise blend weights with ground truth (only known at the SMPL surface). Then we train the canonical mapping network (f_{can}) and the canonical SDF (f_{CSDF}) end-to-end. Then in the final step, we freeze canonical mapping network (f_{can}) and train (f_{CSDF}, f_{def}) together end-to-end. For training the normal prediction network, we train only for points which are near the surface, *i.e.* $|d^*| < \delta$, once the loss of f_{CSDF} has stabilized. We train our network using the following point-wise losses:

$$\mathcal{L}_{\text{SDF}} = ||d^* - d||_2 \qquad \mathcal{L}_{\text{norm}} = 1 - \boldsymbol{n}^T\boldsymbol{n}^* \qquad (7)$$

where d^*, d represent the predicted and ground truth signed distances respectively, and $\boldsymbol{n}^*, \boldsymbol{n}$ denote the predicted and ground truth normals for a point in the posed space. We also added a regularizer for pose dependent displacement field network, given by:

$$\mathcal{L}_{\Delta\bar{\boldsymbol{p}}_{\boldsymbol{\theta}}} = ||\Delta\bar{\boldsymbol{p}}_{\boldsymbol{\theta}}||_2^2 \qquad (8)$$

We generate training data by sampling points on the surface of the ground-truth mesh and randomly displace these points with noise $\sigma = \{0.01, 0.1\}$. Similar to [16], the ground-truth signed distances at these samples are computed by casting randomized rays and checking the parity [24]. We use $K = 24$ number of joints, for fully body meshes and follow [42] for skirt and shirt meshes.

Inference: For generating meshes from our method, we first sample points in space and predict d^*, using Eq. (4). We reconstruct our meshes using marching cubes [30] on the predicted SDF. Our reconstructed meshes are at 256 voxel resolution. Then we predict normals for the vertices on the reconstructed surface, using f_{norm}. Note that unlike [51], we directly query our point in pose space, so there is no need for extra re-posing step.

4. Experiments and Results

Our proposed method can be used to learn non-rigid pose-dependent deformations in human/clothing meshes. We show results of our method on various datasets in Sec. 4.1 and compare with [18, 51] in Sec. 4.3. We also evaluate our model in multiple shape setting in Sec. 4.4.

Datasets: For evaluation of single-shape and clothing model, we use CAPE dataset [31, 44], DFAUST [12, 45], TailorNet [42] and a few human scan sequences with complex clothing that we captured in our stage, which we will refer to as ClothSeq. For multiple subject setup, we evaluate our method on three datasets: 1) MoVi [19, 32], 2) SMPL pose and shape dataset and 3) DFAUST registrations [12, 45] and compare with [36]. For SMPL pose and shape dataset, we use the 890 poses provided in SMPL model and create 9 different body shapes, by changing first 3 principal components of shape-space. For the CAPE and TailorNet, the SMPL parameters are available in the dataset. For others, we fit SMPL [8, 9] to obtain the pose and shape.

We split the dataset, such that the test set contains 2 unseen motion sequences for CAPE and DFAUST. For TailorNet and SMPL pose-shape dataset, we follow the train/test split used in [42]. For MoVi, we follow the train/test split used in [36]. In order to show that Neural-GIF can model complex geometries, loose clothing and topolgies, we test our method on the ClothSeq which is a dataset of complex clothing like jacket, loose shrugs etc.

4.1. Reanimating people in clothing

Modeling people in complex clothing: We use our Neural-GIF, to animate characters in new poses. We generate results for various characters of the ClothSeq and CAPE datasets. We provide results on poses of the test set (volleyball and walking sequences) of the CAPE dataset in Fig. 3 and on the ClothSeq dataset in Fig. 4. Our method generates accurate reconstructions and generalizes to unseen poses, while preserving the structure of identity and generating fine details. Since we directly learn from scans, which are noisy, *e.g.*, near the hands, feet and inner-thigh, the reconstruction around these areas tends to be a bit noisy and unstructured. Modeling such types of clothing using template-based methods is not straightforward due to the open-layered structure of the clothes on top of the body. As a result, a single template cannot explain two extreme deformations of a jacket, (*i.e.*, on one end when a layer of the jacket is close to the body surface and on the other side of the spectrum when the layer is far from the body surface due to some pose like lifting hands in the air). Our method produces large deformations for loose garments like open jackets and shrugs(Fig. 4) as opposed to template-based methods. This is possible because we do not rely on blend weights associated to the underlying body, but instead we predict blend weights for every point in the space as a function of pose, for a given subject. So our network learns to associate each point correctly to joints *e.g.* a point on the surface of jacket near torso will have different association (weight) in the rest pose and slightly different association when the person is lifting hands. Our canonical mapping network learns this pose-dependent association from data.
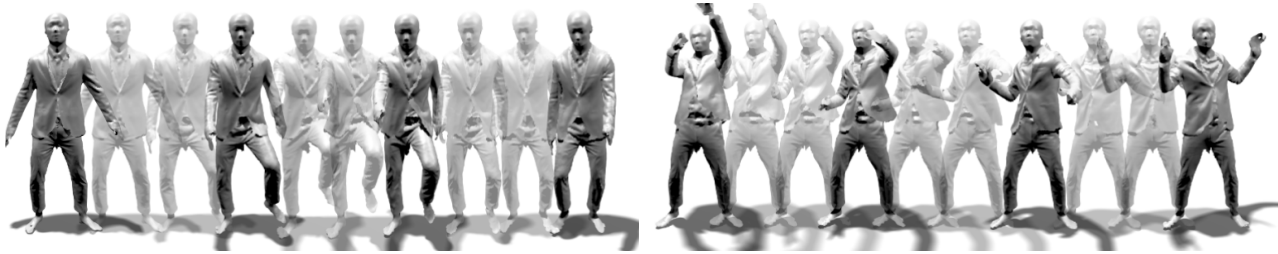
Figure 3. **Re-animating CAPE:** Reposing a character from CAPE dataset [31] using our proposed method. We show results on walking motion sequence (left) and on playing volleyball motion sequence (right).

Figure 4. **Modeling people in complex clothes:** Results of Neural-GIF for subjects from the ClothSeq dataset. Neural-GIF can model well complex cloth geometries, like hoodies (left), puffer jacket (middle) and loose clothes like shrugs (right).

Modeling separate clothing items: Since our method is not limited by a pre-defined template or topology, we can also train Neural-GIF on the clothing meshes of shirts and skirts from the TailorNet dataset, and for completeness, also compare to TailorNet [42]. It is important to note that such a comparison is not totally fair for our method. First, the training meshes are of a fixed resolution dictated by the template used in TailorNet, so our model is upper-bounded by this. Second, registration is noise free because meshes are obtained from physics simulations (ideal case for template-based methods) – this allows the mapping of every shape to a vector, which makes learning significantly easier for TailorNet – correspondence is known for TailorNet whereas is not known for Neural-GIF. Despite this, qualitatively Neural-GIF is comparable to TailorNet [42] as seen in Fig. 5. We can see that our method generalizes to new unseen poses and predicts fine details and deformations. Quantitatively we find that point to surface distance for TailorNet results is slightly better than our method. We produce 11.14 mm and 10.98 mm error for skirt and shirt clothing, while TailorNet's results are 9.82 mm and 7.28 mm respectively. Another source of slight inaccuracy originates from our mesh reconstruction method for open surfaces. Since shirt/skirts are a thin layered surface, we reconstruct the mesh using a small threshold of 0.01mm in marching cubes, which might result in slightly thicker surfaces, and could be addressed by switching to unsigned neural distance fields [16]. This comparison shows that when registration is possible (easy clothing animated with physics simulation) it helps learning.
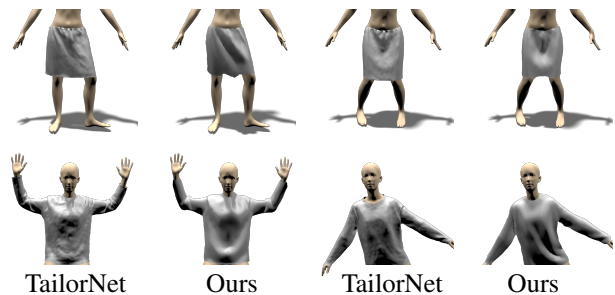

TailorNet   Ours   TailorNet   Ours

Figure 5. **Comparison with TailorNet:** Qualitative comparison on skirt(top) and shirt(bottom) meshes from the TailorNet dataset.

4.2. Modeling soft-tissue dynamics

Neural-GIF models high frequency pose-dependent deformations for various shapes like full clothing, jackets and skirts. We now evaluate how well this representation can model soft-tissue dynamics, which are highly non-linear and high-order pose-dependent deformations. Thus, we evaluate our method on this task using the DFAUST [13] dataset. We show our results on DFAUST [13] in Fig. 6. We evaluate our method on a single subject with 14 sequences, which demonstrate extreme soft-tissue dynamics. In Fig. 6 we provide results for 3 sequences and observe that the network predicts soft-tissue dynamics/deformations. Our normal prediction network also helps in producing more realistic renders. We notice that since deformation in the facial region is not controlled by any keypoint and also the subjects' facial expressions change over time, we tend to produce smooth results around the face.

Table 1. We quantitatively compare the results of our method with NASA [18] and SCANimate [51]. We report point to surface distance (in mm) and IoU and F-Scores(%) for comparison.

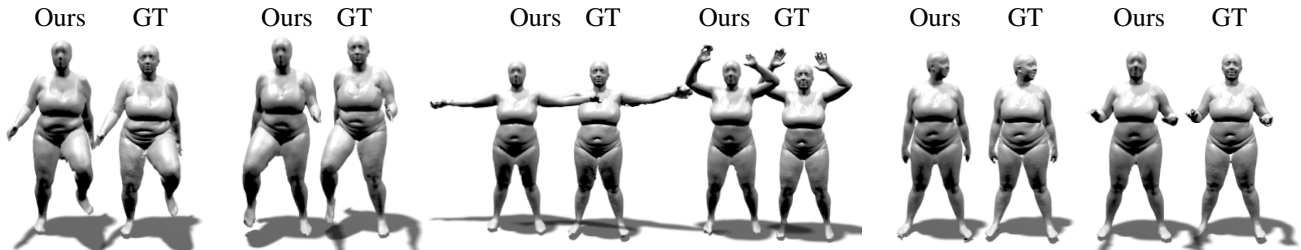| Model	NASA [18]			SCANimate [51]			Ours (Neural-GIF)		
Dataset	Point2Surface ↓	IoU ↑	F-Score ↑	Point2Surface ↓	IoU ↑	F-Score ↑	Point2Surface ↓	IoU ↑	F-Score ↑
CAPE [31]	10.67	0.918	94.32	**5.82**	**0.957**	**98.51**	5.86	0.957	98.53
ClothSeq	23.26	0.780	57.29	7.32	0.953	97.32	**4.73**	**0.967**	**99.15**
DFAUST [13]	10.52	0.939	95.48	3.79	0.971	99.50	**3.21**	**0.972**	**99.56**

Figure 6. **Results on DFAUST [13]:** We show our results on three different motion sequences (running, jiggle on toes, shake hips) from DFAUST and compare with ground-truth DFAUST data and visualize results in (prediction, GT) pairs. Neural-GIF is able to retain pose-dependent soft-tissue deformations present in the ground truth scans.

4.3. Comparisons and Ablation Studies

Comparison with SoTA: For the single shape model, we compare our method with [18, 51] and provide the quantitative evaluations in Table 1. In NASA [18] pose-dependent occupancy is represented using part-based implicit functions. NASA produces good results for human meshes from [13], however it cannot model fine details and complex or loose clothing like jackets. We observe from Fig. 9, that NASA generates rigid and overly smooth results, compared to our model. NASA also suffers from part-based artifacts, such as intersecting body parts in poses like bending elbows or knees. This problem is even more prominent if there are significantly less data points for such poses. Since we learn the whole body shape using a single network our approach does not encounter such issues. From Tab. 1 we observe a significant drop in terms of IoU on ClothSeq when using NASA, which is because the network cannot learn the loose deformations of jackets and hoodies, as one can observe in Fig. 9. Quantitatively, SCANimate [51] performs better than our method on CAPE, as seen in Table 1, first row, but we notice that results of SCANimate are less detailed and have posing artifacts due to LBS, as highlighted in figure 7. We also show significant improvements in terms of the quality of details produced by our method. We accredit this to the combination of the displacement field network, the normal prediction and positional encoding of query point.

Ablation studies: We now evaluate the significance of each module in our proposed method and report the result of ablation in Table 2 and our complete model in Table 1(last column). First, we simply use the Canonical SDF to learn the distance field in the canonical space. We find the blend weights for every point in space (p), by associating the blend weight of nearest point on the SMPL body surface.
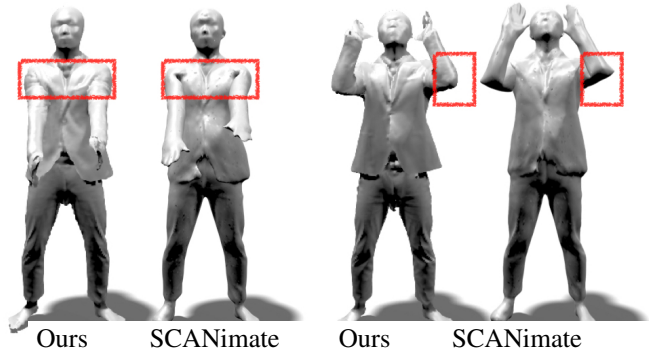

Figure 7. **Comparison with SCANimate:** We compare the results of our method on the CAPE dataset with SCANimate [51]. Our model preserves more details and does not have posing artifacts.

Table 2. **Ablation**: Quantitative evaluation between nearest neighbour based SMPL weights and learned weights from canonical mapping network.

Model	SMPL weights		Canonical mapping network	
Dataset	Point2Surface ↓	IoU ↑	Point2Surface ↓	IoU ↑
CAPE [31]	12.93	0.866	7.93	0.955
ClothSeq	11.89	0.874	9.18	0.963
DFAUST [13]	18.38	0.806	3.28	0.972

We transform p to $\bar{p}$ using Eq. (1) and calculate the SDF using Eq. (4). In Fig. 8(left), we observe that if we simply use nearest point based blend weights, we obtain wrong reconstructions (space between the legs). We observe that for this case (Table 2, col.1), the IoU is relatively low as compared to the other two versions of our model and also lower than NASA(see Table 1) on DFAUST. This is because of incorrect body part associations, as shown in Fig. 8 and explained above. This problem is most prominent, near the inner thighs and especially when both the legs are not per-

$f_{\mathrm{CSDF}}+$ SMPL weights $f_{\mathrm{CSDF}}+f_{\mathrm{can}}$ Ours (complete)
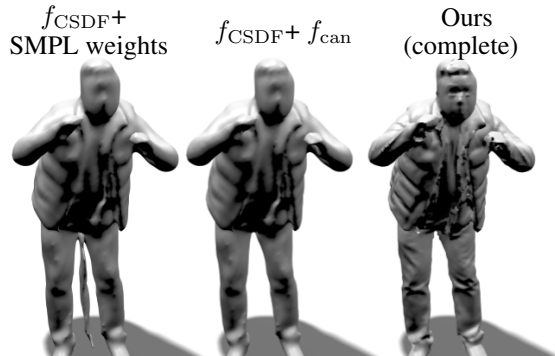
Figure 8. **Ablation:** Canonical SDF+SMPL weights results in wrong association and hence result in wrong occupancy (left). Our canonical mapping network solves this problem by learning blend weights from data, but produces smooth results (middle). Finally displacement field network and normal prediction network helps in generating small pose-dependent details (right).

forming a symmetrical motion. The test split in DFAUST consists of jumping, running etc, the points between the legs randomly change their association between left and right legs and are mapped to a point in the canonical space, which may lie inside the canonical shape, for a given pose.

We alleviate this problem, with our canonical mapping network to predict these blend weights (Table 2, col.2). The network learns correct associations based on data. Learning the SDF in the canonical space, simplifies the learning process, as the network learns a strong prior for the shape of the subject. We observe from Fig. 8(middle), that just using Canonical SDF + Canonical mapping network, does not produce fine pose-dependent deformations. Hence we introduce a displacement field network, to learn a pose-dependent displacement field in the canonical space for small non-rigid deformations and a normal prediction model. We observe that having a separate model to incorporate such fine details and training the network in a multi-step regime (Sec. 3.5), helps to learn these deformations.

4.4. Multiple shape model

As mentioned in section 3.4, we extend our formulation to incorporate multiple body shapes in one model. We compare with concurrent method LEAP [36] and show the quantitative evaluation in Table 3. We model body shape variation as a continuous displacement field in canonical space. Even with a simple formulation, we perform comparably or better than LEAP. We also observe that our training is more stable than training LEAP.

5. Conclusion

We introduced Neural-GIF, a novel model to learn articulation and pose-dependent deformation for humans in complex clothing using an implicit 3D surface representation. Neural-GIF admits end-to-end learning directly from

Table 3. For multiple shape model we quantitatively compare the results of our method with LEAP [36].

Model Dataset	LEAP [36]		Ours (Neural-GIF)	
	Point2Surface ↓	IoU ↑	Point2Surface ↓	IoU ↑
DFAUST [13]	3.42	0.958	**3.35**	**0.963**
MoVi [19]	**3.19**	**0.969**	3.20	0.969
SMPL	3.26	0.968	**3.18**	**0.971**

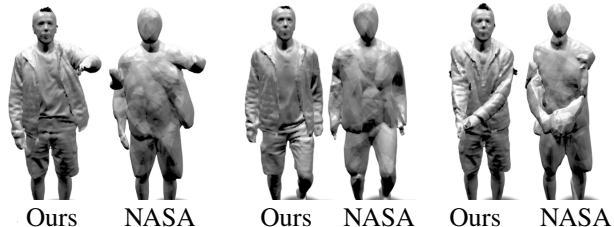Ours   NASA   Ours   NASA   Ours   NASA

Figure 9. **Comparison with NASA:** We compare the results of our method (bottom) on the ClothSeq dataset with NASA [18] (top). Our model reconstructs complex clothing without any part-based artifacts and generates realistic pose-dependent deformations.

scans provided with their corresponding SMPL pose and shape parameters. The key idea of Neural-GIF is to express shape as displacement and articulation deformations of a canonical shape, a concept that has been widely used for parametric meshes, which we generalize here to neural implicit function learning. Neural-GIF can accurately model complex geometries of arbitrary topology and resolution, because our model does not require a pre-defined template, or non-rigid registration of a template to scans.

We show significant improvements from prior work [18, 51], in terms of robustness, ability to model complex clothing styles and retaining fine pose-dependent details. We believe that the use of generalized implicit functions with our canonical mapping and displacement field networks help the network to more effectively factor out articulation from non-rigid components. We further extend our model to multiple shape setting and show comparable performance to concurrent work LEAP [36].

We believe that Neural Generalized Implicit Functions open several interesting research directions. Since currently we have a clothing-specific model, it will be useful to extend this approach such that it can animate multiple clothing using the same model. It would be also valuable to learn temporal correspondences implicitly during learning. Finally, since Neural-GIF produces signed distance fields, we want to leverage them for fast collision and contact computation in human-object and scene interactions.

References

[1] Neural-GIF: Neural Generalized Implicit Functions for Animating People in Clothing. https://virtualhumans.mpi-inf.mpg.de/neuralgif/, 2021. 1

[2] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019. 1

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[4] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 2

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, July 2005. 2

[6] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. CLOTH3D: clothed 3d humans. volume abs/1912.02792, 2019. 2

[7] Hugo Bertiche, Meysam Madadi, Emilio Tylson, and Sergio Escalera. Deepsd: Automatic deep skinning and pose space deformation for 3d garment animation, 2021. 3

[8] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 2, 5

[9] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020. 3, 5

[10] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 2

[11] 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 *European conference on computer vision*, 2016. 1

[12] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 5

[13] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 6, 7, 8

[14] Aljaž Božič, Pablo Palafox, Michael Zollhöfer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deforma-

[15] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 2, 3

[16] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020. 3, 5, 6

[17] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 3

[18] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa: Neural articulated shape approximation. In *The European Conference on Computer Vision (ECCV)*, August 2020. 2, 3, 5, 7, 8

[19] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. Movi: A large multi-purpose human motion and video dataset. *PLOS ONE*, 16(6):e0253157, Jun 2021. 5, 8

[20] P. Guan, L. Reiss, D. Hirshberg, A. Weiss, and M. J. Black. DRAPE: DRessing Any PErson. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 31(4):35:1–35:10, July 2012. 2

[21] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 2

[22] Marc Habermann, Weipeng Xu, , Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1

[23] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 3

[24] Alec Jacobson, Ladislav Kavan, and Olga Sorkine. Robust inside-outside segmentation using generalized winding numbers. *ACM Trans. Graph.*, 32(4), 2013. 5

[25] Timothy Jeruzalski, David I. W. Levin, Alec Jacobson, Paul Lalonde, Mohammad Norouzi, and Andrea Tagliasacchi. Nilbs: Neural inverse linear blend skinning, 2020. 3

[26] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. *arXiv preprint arXiv:2004.00214*, 2020. 2

[27] Z. Laehner, D. Cremers, and T. Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *European Conference on Computer Vision (ECCV)*, September 2018. 2, 3

tion graphs for globally-consistent non-rigid reconstruction, 2020. 3

[28] Lijuan Liu, Youyi Zheng, Di Tang, Yi Yuan, Changjie Fan, and Kun Zhou. Neuroskinning: Automatic skin binding for production characters with deep graph networks. *ACM Trans. Graph.*, 38(4), July 2019. 3

[29] 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, 4

[30] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *COMPUTER GRAPHICS*, 21(4):163–169, 1987. 5

[31] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael Black. Learning to dress 3d people in generative clothing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1, 2, 5, 6, 7

[32] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 5

[33] Takashi Matsuyama, Shohei Nobuhara, Takeshi Takai, and Tony Tung. *3D Video and Its Applications*. Springer, 2012. 2

[34] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[35] Eder Miguel, Derek Bradley, Bernhard Thomaszewski, Bernd Bickel, Wojciech Matusik, Miguel A. Otaduy, and Steve Marschner. Data-driven estimation of cloth simulation models. *Comput. Graph. Forum*, 31(2):519–528, 2012. 2

[36] 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. 2, 3, 5, 8

[37] 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 *ECCV*, 2020. 3, 5

[38] Matthias Müller and Nuttapong Chentanez. Wrinkle Meshes. In MZoran Popovic and Miguel Otaduy, editors, *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation*. The Eurographics Association, 2010. 2

[39] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pages 598–613, 2020. 2

[40] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[41] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Deformable neural radiance fields. *arXiv preprint arXiv:2011.12948*, 2020. 3

[42] Chaitanya Patel, Zhouyincheng Liao, and Gerard Pons-Moll. The virtual tailor: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2020. 1, 2, 3, 5, 6

[43] 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)*, pages 10975–10985, 2019. 1

[44] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4), 2017. 2, 3, 5

[45] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: a model of dynamic human shape in motion. *ACM Transactions on Graphics*, 34:120, 2015. 1, 2, 3, 5

[46] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes, 2020. 3

[47] Damien Rohmer, Tiberiu Popa, Marie-Paule Cani, Stefanie Hahmann, and Sheffer Alla. Animation Wrinkling: Augmenting Coarse Cloth Simulations with Realistic-Looking Wrinkles. *ACM Transactions on Graphics*, 29(5):0, Dec. 2010. Proceedings of ACM SIGGRAPH ASIA. 2

[48] Bodo Rosenhahn, Uwe Kersting, Katie Powell, Reinhard Klette, Gisela Klette, and Hans-Peter Seidel. A system for articulated tracking incorporating a clothing model, 2007. 2

[49] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[50] 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. 2

[51] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2, 3, 5, 7, 8

[52] Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-Based Animation of Clothing for Virtual Try-On. *Computer Graphics Forum (Proc. Eurographics)*, 2019. 1, 2

[53] Stan Sclaroff and Alex Pentland. Generalized implicit functions for computer graphics. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '91, page 247–250, New York, NY, USA, 1991. Association for Computing Machinery. 2, 3

[54] Carsten Stoll, Juergen Gall, Edilson de Aguiar, Sebastian Thrun, and Christian Theobalt. Video-based reconstruction of animatable human characters. 29(6), Dec. 2010. 2

[55] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 2

[56] Tony Tung, Shohei Nobuhara, and Takashi Matsuyama. Simultaneous super-resolution and 3d video using graph-cuts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[57] Tony Tung, Shohei Nobuhara, and Takashi Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *IEEE 12th International Conference on Computer Vision, ICCV*, pages 1709–1716, Kyoto, Japan, Sept. 2009. 2

[58] Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. Fully Convolutional Graph Neural Networks for Parametric Virtual Try-On. *Computer Graphics Forum (Proc. SCA)*, 2020. 2

[59] Huamin Wang, Ravi Ramamoorthi, and James F. O'Brien. Data-driven elastic models for cloth: Modeling and measurement. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 30(4):71:1–11, July 2011. 2

[60] Tuanfeng Y. Wang, Duygu Ceylan, Jovan Popovic, and Niloy J. Mitra. Learning a shared shape space for multimodal garment design. *ACM Trans. Graph.*, 37(6):1:1–1:14, 2018. 2

[61] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, 2019. 1

[62] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render and compare. In *International Conference on Computer Vision*, 2019. 1, 2

[63] Ze Yang, Shenlong Wang, Siva Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *CVPR*, 2021. 3

[64] T Yenamandra, A Tewari, F Bernard, HP Seidel, M Elgharib, D Cremers, and C Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3

[65] Meng Zhang, Tuanfeng Wang, Duygu Ceylan, and Niloy J Mitra. Deep detail enhancement for any garment. *arXiv preprint arXiv:2008.04367*, 2020. 2