NSF: Neural Surface Fields for Human Modeling from Monocular Depth

Yuxuan Xue[1,2,*], Bharat Lal Bhatnagar[1,2,3,4,*], Riccardo Marin[1,2], Nikolaos Sarafianos[4], Yuanlu Xu[4], Gerard Pons-Moll[1,2,3,†], Tony Tung[4,†],

[1] Tübingen AI Center [2] University of Tübingen

[3] Max Planck Institute for Informatics [4] Meta Reality Labs Research

[†] Project Lead

{yuxuan.xue, riccardo.marin, gerard.pons-moll}@uni-tuebingen.de

{bharatbhatnagar, nsarafianos, yuanluxu, tony.tung}@meta.com

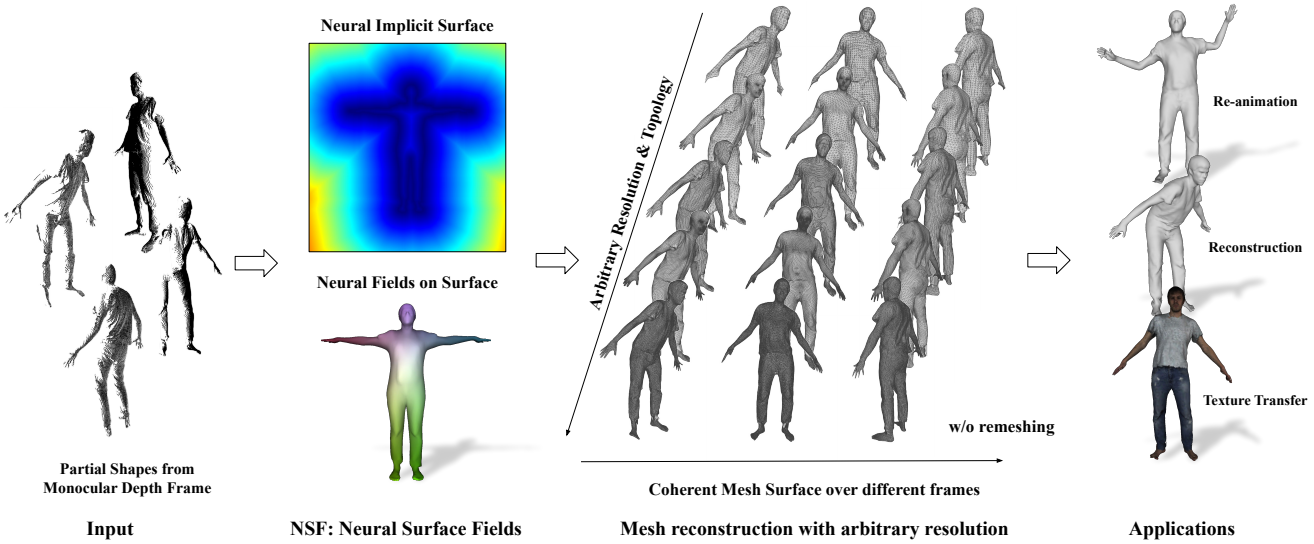

Figure 1. Given a sequence of monocular depth frames of a person, NSF learns a detailed clothed body model of the person. The clothed body model is controllable and can be used for reconstruction and re-animation at arbitrary mesh resolution while maintaining the coherency whithout retraining.

Abstract

Obtaining personalized 3D animatable avatars from a monocular camera has several real world applications in gaming, virtual try-on, animation, and VR/XR, etc. However, it is very challenging to model dynamic and fine-grained clothing deformations from such sparse data. Existing methods for modeling 3D humans from depth data have limitations in terms of computational efficiency, mesh coherency, and flexibility in resolution and topology. For instance, reconstructing shapes using implicit functions and extracting explicit meshes per frame is computationally expensive and cannot ensure coherent meshes across frames. Moreover, predicting per-vertex deformations on a pre-designed human template with a discrete surface lacks flexibility in resolution and topology. To overcome these limitations, we propose a novel method 'NSF : Neural Surface Fields' for modeling 3D clothed humans from monocular depth. NSF defines a neural field solely on the base surface which models a continuous and flexible displacement field. NSF can be adapted to the base surface with different resolution and topology without retraining at inference time. Compared to existing approaches, our method eliminates the expensive per-frame surface extraction while maintaining mesh coherency, and is capable of reconstructing meshes with arbitrary resolution without retraining. To foster research in this direction, we release our code in project page at: https://yuxuan-xue.com/nsf.

* denotes equal contribution

1. Introduction

Human modeling is an active and challenging field of research that has applications in Computer Vision and Graphics. Recent advancements in data acquisition techniques [21, 60, 61, 74, 75, 77, 48] have opened new opportunities for capturing and digitising human appearance. Building digital avatars has found applications in behavioural studies [12, 16, 18, 22, 55, 71, 72] and generative modelling [23, 33, 35, 58]. Our goal is to build body model which is controllable i.e., animatable with different poses, and detailed i.e. it should faithfully produce details such as garments wrinkles under different poses.

In recent years, researchers have looked into learning clothed human models from full sequences of 4D scans [8, 34, 39, 41, 62, 67]. 4D scans provide rich information about the subject appearance, but they also require exclusive technology, pre-processing, and expert intervention at times, which makes this difficult to scale. A more user friendly line relies on the input with monocular depth from devices such as Kinects [6, 13, 28, 78, 79]. Such data is easier to obtain and already supported by consumer-grade devices. But this flexibility comes at the cost of additional sensor noise, thus complicating the learning process.

To mitigate the noise in input data, parametric models such as SMPL [36] and its successors [1, 4, 53, 76, 3], can provide a good statistical prior for capturing pose and the overall shape of the person. Also, relying on a template naturally supports information transfer across subjects and poses. However, designing a pipeline around a specific template restricts the expressivity of the model, which makes the methods less flexible (e.g., limited to tight garments). A common representation to relax the topology constraints is point clouds [34, 39, 41, 81]. Recently, point based neural implicit representations [8, 13, 62, 67, 69, 2] demonstrated incredible expressive power. But many real applications (e.g., animation, texture transfer) require a 3D mesh. Hence, these approaches require running costly algorithms [37, 27] to reconstruct a supporting surface. Extracting a surface for every frame causes a computational burden and also results in inconsistent triangulations, which further complicate downstream tasks. Some works [6, 28] address this issue by predicting displacements on SMPL vertices for modeling clothed humans. While these methods yield coherent mesh reconstruction, they are constrained by the resolution and topology of SMPL template.

We pose ourselves the following goal: starting only from a set of partial shapes from monocular depth frames, can we learn a clothed body model that is *flexible* and *coherent* across different frames, with a *limited computational cost for surface extraction*?

To this end, we propose *NSF : Neural Surface Fields*; a neural field defined continuously all over the surface. Given a canonical shape, represented with an implicit function, we use NSF to define a continuous field over the surface, capable of modeling detailed deformations. Using NSF, we can reconstruct a *coherent mesh* in the canonical space at any resolution with just one run of surface extraction algorithms, and share it across all the different poses. This formulation avoids per-frame surface extraction which is $\sim$ 40x and $\sim$ 180x faster compared to point-based works [34, 39, 41, 81] using Poisson reconstruction and implicit-based works [8, 13, 62, 67, 69] using marching cube at similar resolution, respectively. After training, NSF can be adapted to *arbitrary resolutions* at inference time, depending on the application. This step is possible since NSF is continuously defined all over the surface, and hence it is able to support any discretization. Compared to other feature representations, NSF is more compact, saving 97.4% of memory compared to volumetric representation and 86.0% compared to triplane features at 128^3 resolution.

We validate our self-supervised approach on several datasets [6, 28, 32, 40, 56], showing better performance than competitors, even when some of them requires subject-specific training [6, 13, 41, 49, 50, 69]. We show the practical benefits of NSF in shape reconstruction, animation, and texture transfer application, with the flexibility and the coherency that is not attainable for prior works [6, 13].

In summary, our contributions can be summarized as:

- We propose *NSF : Neural Surface Fields*; a continuous neural field defined over the surface in a canonical space which is compact, efficient, and supports arbitrary mesh discretizations without retraining.

- We propose a method to learn an animatable human avatar from a monocular depth sequence; NSF let us recover detailed shape information from monocular depth frames. Our self-supervised approach handles subjects with different clothing geometries and textures. To the best of our knowledge, NSF is the first work in avatarization which directly output mesh at arbitrary resolution while maintaining the coherency across different poses.

2. Related Work

Human Capture. Clothed human reconstruction is a rapidly evolving field of research that aims to create realistic and detailed digital models of humans. Recent work [19, 20, 24, 60, 61, 75, 74, 83] can reconstruct humans from a single RGB image but are not as accurate. Methods such as KinectFusion [47] and DynamicFusion [46] fuse depth measurements over time to create a complete and accurate model. While these are general and not restricted to humans, BodyFusion [78] and DoubleFusion [79] incorporate priors on human motion and shape, fusing partial depths in real-time to obtain improved reconstruction.

However, these methods are complicated to setup and require expert intervention. Moreover, their code is unavailable. With the advent of deep learning methods, data-driven methods such as IF-Nets [9], reconstruct humans by learning a prior from a large dataset. IP-Net [2] further fits a parametric model to the implicit reconstruction to make the mesh controllable. These approaches only capture static humans and do not capture the pose dependent deformations, thus lacking realism.

Implicit Neural Avatar. In the last few years, outstanding results produced by Neural Radiance Field (NeRF) [44] have motivated scholars to model the clothed human as implicit neural representations. There's a plethora of NeRF-based approaches for humans modeling that provide animatable avatars starting from monocular RGB videos [15, 14, 25, 54, 63, 70, 82]. Apart from constructing the human model using RGB images, a common and straightforward approach involves learning the implicit neural avatar from geometric data, such as scans [7, 8, 11, 43, 67, 69, 85, 66]. Furthermore, PINA [13] models an implicit personalized avatar using monocular depth sequences, which share the same input as our work. However, it is important to note that these implicit-based methods are *subject-specific* and are unable to model multiple subjects simultaneously. Furthermore, these methods that rely on implicit representations utilize neural networks to parameterize the shape, and cannot directly provide explicit meshes as output. In order to obtain a mesh representation, an extensive computation of marching cubes is performed for *each frame*, resulting in computationally expensive operations. Moreover, the extracted surface using marching cubes lacks *coherence* across different frames. This lack of coherency leads to the loss of natural correspondence and poses additional hindrances in applying these methods to downstream tasks, e.g. texture transfer between the input and the learned shape.

Explicit Parameterized Avatars. SMPL [36] is a popular parametric human model. However, it only models the naked body shape and pose, and lacks details. Hence, several extensions have been proposed to add further details like hands [59], face [53], soft-tissues [57] and clothing [52, 56, 57, 80, 4]. Many works model deformations [6, 28, 39, 41, 38, 80, 81, 1], by fitting SMPL model and adding cloth wrinkles as displacement on top of the coarse shape. Although they reconstruct coherent shapes, they are often limited by the resolution and topology of the SMPL template, making them less flexible compared to implicit-based methods. To overcome this limitation, Lin *et al.* [34] proposed to learn the fusion shape using implicit occupancy network, which is not constrained by the SMPL topology and can represent loose garments like skirts. However, this approach relies on complete scans and registered mesh data to provide ground-truth occupancy labels. Moreover, these point-based works [34, 39, 41] need to perform Poisson Reconstruction at each frame to obtain the mesh. In contrast, our approach fuses monocular raw depth inputs into a canonical space to obtain a coarse, pose-independent base shape without any supervision, which is difficult to obtain from partial shape data. We then learn pose-dependent neural surface fields (Sec. 3) on top of the coarse shapes, which allow us to obtain detailed shapes at arbitrary resolutions. In summary, our approach offers flexibility and efficiency in generating coherent meshes, and eliminates the need for Marching Cubes or Poisson Reconstruction at each frame (Sec. 4.4).

3. NSF: Neural Surface Fields

Neural Fields. A neural field is a field parametrized by a neural network [73]:

$$f_\phi : \mathbb{R}^m \rightarrow \mathbb{R}^n, \tag{1}$$

where ϕ are the learnable parameters. Neural field defined in Euclidean space $\mathbb{R}^3$ has been widely-used to represent various geometries like distance [51], occupancy [42], and radiance [44] functions, correspondences [2], contacts [26, 5, 84], parametric body models [3], and so on.

Neural Surface Fields. When a field carries information about an object that occupies a limited volume bounded by a 2D surface $\mathcal{S}$, we know in advance that much region of the space will not be ever queried, causing a waste of computational and memory resources [9, 2, 3]. Following this intuition, we are interested to define the field only on the 2D surface $\mathcal{S}^2$:

$$f_\phi : \mathcal{S}^2 \subset \mathbb{R}^3 \rightarrow \mathbb{R}^n. \tag{2}$$

We call this representation *Neural Surface Fields (NSF)*. Recent work [29] defines the neural field with the eigenfunction of the Laplace-Beltrami Operator on the surface, and hence are defined just for a specific discretization of the geometry. Instead, our approach is more general and produces a continuous field independent of the underlying discretization of the object.

Embedding the neural fields on a surface is advantageous due to the ability to combine properties with mesh surface coherency and connectivity as shown in Fig. 2. In our work, we leverage NSF to learn a continuous deformation field which models the detailed clothing deformations on the surface of the coarse clothed human shape (Sec. 4.2).

4. NSF for Human Modelling

In this section we show the advantages of NSF by incorporating it into an avatarization method. Before diving into the method details, we will state our goal, define method's input, and provide a general overview.

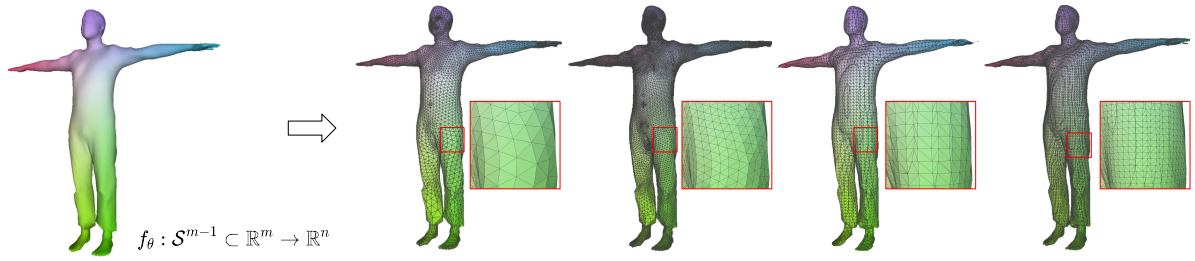

Figure 2: We show an example of NSF decoding to surface color $\in \mathbb{R}^3$. On the right of arrow shows that NSF can be queried with the surface with arbitrary resolution or topology without retraining.

Input. Let $s = \{1, ...N\}$ be the set of subjects. For each subject, our method takes as input a sequence of monocular depth point clouds, $\mathcal{X}^s = \{\mathbf{X}_1^s, ...\mathbf{X}_{T_s}^s\}$. Each $\mathbf{X}_t^s$ is a set of unordered points $\{\mathbf{x}_j^{s,t}\}_{j=1}^{L_{s,t}}$ where $L_{s,t}$ represents the number of points in the monocular point cloud at time t. Also, for subject sequence we take as input the corresponding 3D poses $\theta^s = \{\theta_1^s, ...\theta_{T_s}^s\}$.

Output. Our goal is to learn subject-specific body models, $\mathcal{M} = \{M^1, ...M^N\}$. Each model $M^s(\mathbf{p}, \theta)$ can transform points $\mathbf{p} \in \mathbb{R}^3$ from a neutral pose in canonical space to the target pose θ, taking the shape and clothing of the subject into account. Our models are complete, detailed, and contain pose dependent garment deformations of the subject.

Overview. We kindly ask readers to refer Fig. 3 for an overview of our method. To learn the body model of each subject, (A) we unpose the input point clouds (Sec. 4.1) to a neutral pose using inverse skinning, and (B) we fuse them to learn an implicit (SDF) *canonical shape* $\mathcal{B}^s$ (Sec. 4.1). Our canonical shape is continuous, and the fusion of different depths averages out fine-grained details generated by the subject poses. On top of our canonical shape, (C) we train NSF (Sec. 4.2), which predicts the pose-dependent deformation for each point on the continuous canonical surface, (D) recovering the cloth deformation for a specific pose of the subject (Sec. 4.2). Finally, (E) we use LBS to pose the human model (Sec. 4.3). The method is optimized using a cycle-consistency loss between the input point cloud and our predicted shape. For simplicity we drop s from subsequent notation and explain our method for a single subject. We will reintroduce s for parts of the manuscript dealing with multiple subjects.

4.1. Fusion Shape from Monocular Depth

Canonicalization. To build our person-specific canonical shape, we unpose every $\mathbf{X}_t$ input point cloud to a neutral pose. The corresponding canonical points $\mathbf{X}_t^c$ for input points can be found using iterative root finding [8, 31]:

$$\underset{\mathbf{X}_t^c, w}{\arg\min} \sum_{t=1}^{T} \left(\left(\sum_{i=1}^{K} w(\mathbf{X}_t^c)_i \cdot \mathbf{T}_i(\theta_t) \right) \mathbf{X}_t^c - \mathbf{X}_t \right). \quad (3)$$

where K is the number of joints, and $w(\cdot)_i$ and $\mathbf{T}_i$ are the skinning weights and joint transformation for joint i respectively. We utilize the iterative root finding in canonicalization together with the pre-diffused SMPL skinning field in FiTE [34] to avoid ambiguous solutions. We unpose all input observation $\mathcal{X} = \{\mathbf{X}_t\}_{t=1}^T$ into canonical partial shapes $\mathcal{X}^c = \{\mathbf{X}_t^c\}_{t=1}^T$.

Implicit Fusion Shape. Since the inverse skinning does not account for pose-dependent deformations operates at a human level, the point cloud $\mathbf{X}_t^c$ resulting from our canonicalization process still contains non-rigid deformation specific to the subject poses. To remove the influence of single poses and obtain a coarse canonical shape $\mathcal{B}$, our idea is to fuse every $\{\mathbf{X}_t^c\}_{t=0}^T$ by learning an implicit surface in the canonical space. Concretely, we represent $\mathcal{B}^s$ as an implicit SDF in [51], composed by a neural network $f^{\text{shape}}(\cdot|\phi^{\text{shape}})$ parameterised by parameters ϕ^{shape}, that takes as an input a subject specific latent code $\mathbf{h}^s \in \mathbb{R}^{256}$ and a query point $\mathbf{x} \in \mathbb{R}^3$, to predict an SDF value. The subject-specific latent codes $\mathcal{H} = \{\mathbf{h}^s\}_{s=1}^N$, and the decoder parameters ϕ^{shape}, are optimised with the self-supervised objective [17] below:

$$E^{\text{shape}}(\phi^{\text{shape}}, \mathcal{H}) = E_{\text{geo}} + \lambda_1 E_{\text{eik}} \quad (4)$$

$$E_{\text{geo}}(\phi^{\text{shape}}, \mathcal{H}) = \sum_{s=1}^{N} \sum_{t=1}^{T^s} \sum_{i=1}^{L_{s,t}} \Big(|f^{\text{shape}}(\mathbf{x}_i^c, \mathbf{h}^s|\phi^{\text{shape}})| + \lambda_3 |\nabla_{\mathbf{x}} f^{\text{shape}}(\mathbf{x}_i^c, \mathbf{h}^s|\phi^{\text{shape}}) - \mathbf{n}_i^c|_2 \Big), \quad (5)$$

where $\mathbf{n}_i^c$ is the normal obtained by canonicalising the normal $\mathbf{n}_i$, along with the point $\mathbf{x}_i$ as described in Eq. 3, and $\nabla_{\mathbf{x}}$ denotes the spatial derivative. We compute the normal $\mathbf{n}_i$ on the point cloud using [47]. The term $E_{\text{eik}}(\cdot)$ [17] enforces that the SDF prediction on the canonical surface should be zero and its derivative, i.e. normal direction,

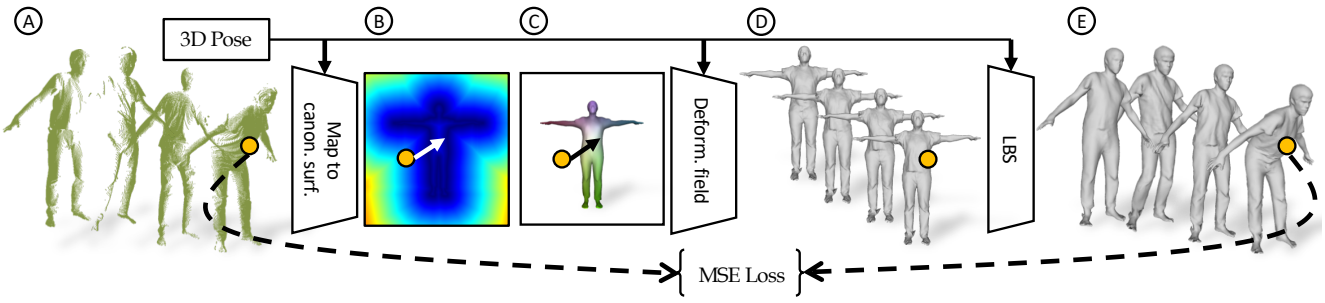

Figure 3: We propose a method to learn animatable body models of people from monocular depth point clouds and 3D poses (A). We learn an implicit canonical shape of a person (B) by fusing the partial point clouds. To get fine details, we learn pose-dependent deformations as a continuous field on the surface of the fusion shape (C), using our *neural surface fields*. By predicting deformations in canonical pose (D), we pose our 3D reconstructions using simple LBS (E). Our approach can be trained with self-supervision.

should match the canonicalised normal:

$$E_{\text{eik}}(\phi^{\text{shape}}, \mathcal{H}) = \sum_{s=1}^{N} \sum_{t=1}^{T^s} \sum_{i=1}^{L_{s,t}} \left(|\nabla_{\mathbf{x}} f^{\text{shape}}(\mathbf{x}_i^c, \mathbf{h}^s | \phi^{\text{shape}})|_2 - 1 \right)^2. \tag{6}$$

Insights. Our objective $E^{shape}(\phi^{shape}, \mathcal{H})$ allows us to fuse all partial canonical frames into a single continuous shape for each subject, averaging out the pose-dependent artefacts. The subject-specific geometry of the canonical shape can be encoded in their respective latent codes $\mathbf{h}^s$, whereas the decoder can freely learn common information across subjects.

4.2. NSF for Pose-Dependent Deformation

Neural Surface Deformation Field. In the previous Section we described how to learn a pose-independent fusion shape by fusing input observations. But to faithfully reproduce the detailed 3D shape of a person we need to model fine-grained pose-dependent deformations. Leveraging the NSF introduced in Sec. 3, we define a deformation field on the top of the fusion shape surface $\mathcal{B}^s$:

$$f_\phi : \mathcal{S}^2 \subset \mathbb{R}^3 \rightarrow \mathbb{R}^3, \tag{7}$$

where points on the surface $\mathcal{S}^2$ are mapped to their corresponding pose-dependent displacements $\mathbb{R}^3$ in the canonical space. Similar to our fusion shape, our deformation fields are also parameterized by a combination of subject-specific latent codes $\mathcal{F} = \{\mathbf{F}^s\}_{s=1}^N$, and a pose conditioned decoder network $f^{\text{pose}}(\cdot|\phi^{\text{pose}})$. More specifically, the deformed points for the subject s is computed as:

$$\mathbf{X}^p = \mathbf{X}^c + f^{\text{pose}}(\mathbf{F}^s(\mathbf{X}^c), \theta | \phi^{\text{pose}}), \tag{8}$$

where $\mathbf{F}^s(\mathbf{x}^c)$ denotes the latent feature queried at point $\mathbf{x}^c$ for subject s and θ denotes the pose feature encoded by a MLP. Our key idea is to learn a NSF for deformation directly and solely on the surface of the implicit fusion shape $\mathcal{B}^s \subset \mathbb{R}^3$ for each subject. This requires addressing two key challenges: *how to learn features* $\mathbf{F}^s(\cdot)$ *on the surface*? and *how to handle off-surface query points for prediction*?.

Feature Learning On Surface. Volumetric and pixel-aligned implicit feature learning methods [2, 9, 60, 61] learn features at regular grid locations and use bi-/tri-linear interpolation to compute features at intermediate points. We devise a similar strategy to learn features on a surface. We first discretize the implicit fusion shape $\mathcal{B}^s$ by Marching Cubes [37] to extract an explicit surface. Moreover, if the garments can be represented by SMPL [36] topology, we fit the SMPL+D model by minimizing the SDF value of SMPL vertices. The same explicit mesh topology allows us to quickly initialize feature space across different subjects. We use the vertices $(5,000 \sim 7,000)$ on this surface to form the feature basis location of our surface. The features are learnt via an auto-decoder during training. The feature $\mathbf{F}^s(\mathbf{x}^c)$ at arbitrary surface point $\mathbf{x}^c \in \mathcal{B}^s$ is obtained using barycentric interpolation between three nearest neighbours among the sampled basis points. Our feature learning on surface is compact and unlike the 1D vectors retains 3D spatial arrangement. In addition, it is memory-efficient, whereas volumetric latent features [9, 2, 10] at 128 resolution require learning $128^3 \sim$ 2mil. features, while we only need to learn about 7k. features using a neural surface space. Our experiments demonstrate that learning a deformation field on a surface produces better results than volumetric and other competing feature learning approaches with significantly lower number of features.

Projecting Off-surface Points Onto Surface. Feature learning on surface is quite straightforward and intuitive as described above. But it requires the query point $\mathbf{x}^c$ to lie on the surface $\mathcal{B}^s$ as the NSF is not even defined outside in $\mathbb{R}^3$. This is challenging because the canonical point $\mathbf{x}^c$ obtained by canonicalising the input observation $\mathbf{x}$ (Eq. 3) is pose-dependent and does not lie on the surface. To this end we use a simple method to project off-surface canonical point to $\mathcal{B}^s$ [10, 65]. We use our pre-trained auto-decoder in Sec. 4.1 to obtain the SDF corresponding to the canonical

point, and the gradient of this SDF gives us the normal direction perpendicular to the surface. We can use this to find the canonical surface point $\mathbf{x}^{cc}$ corresponding to $\mathbf{x}^c$.

$$\mathbf{x}^{cc} = \mathbf{x}^c + f^{\text{shape}}(\mathbf{x}^c, \mathbf{h}^s | \phi^{\text{shape}}) \nabla_{\mathbf{x}_c} f^{\text{shape}}(\mathbf{x}^c, \mathbf{h}^s | \phi^{\text{shape}}). \tag{9}$$

With this surface projection we can obtain the correspondence $\mathbf{x}^{cc}$ on the fusion shape of each pose-dependent canonical point $\mathbf{x}^c$. Afterwards, we can lift the neural surface feature from $\mathbf{x}^{cc}$, $\mathbf{F}^s(\mathbf{x}^c) \leftarrow \mathbf{F}^s(\mathbf{x}^{cc})$.

4.3. Self-supervised Cycle Consistency

Reposing via Skinnning. Once we obtain the pose-dependent deformation on top of the fusion shape in 8, we use standard linear blend skinning [30], to repose points:

$$\mathbf{X}^{pp} = \left(\sum_{i=1}^{K} w_i(\mathbf{X}^p) \mathbf{T}_i(\theta) \right) \mathbf{X}^p, \tag{10}$$

where $\mathbf{X}^p = \{\mathbf{x}_i^p\}_{i=1}^{L_t}$ is the NSF predicted pose-dependent canonical points and $\mathbf{X}^{pp} = \{\mathbf{x}_i^{pp}\}_{i=1}^{L_t}$ is the reposed pose-dependent points. Note that $\mathbf{X}^{pp}$ can be considered as the reconstruction of input observation $\mathbf{X}_t$.

Self-supervised Learning. The NSF, namely subject-specific surface features $\mathcal{F} = \{\mathbf{F}^s\}_{s=1}^{N}$ together with the pose-conditioned decoder network $f^{\text{pose}}(\cdot|\phi^{\text{pose}})$ can be trained end-to-end by ensuring that our posed reconstruction $\mathbf{X}^{pp}$ matches the input point cloud $\mathbf{X}_t$. This can be formulated as the following self-supervised objective:

$$E^{\text{pose}}(\phi^{\text{pose}}, \mathcal{F}) = \sum_{s=1}^{N} \sum_{t=1}^{T^s} \sum_{i=1}^{L_{s,t}} \Big(|\mathbf{x}_i - \mathbf{x}_i^{pp}|_2 + |\mathbf{n}_i - \mathbf{n}_i^{pp}|_2 + d^{\text{CD}}(\mathbf{x}_i, \mathbf{x}_i^{pp}) + E_{\text{reg}}^{\text{pose}} \Big), \tag{11}$$

$$E_{\text{reg}}^{\text{pose}} = |\mathbf{x}_i^p - \mathbf{x}_i^c|_2 + |\mathbf{F}^s(\mathbf{x}_i^c)|_2 + EDR(\mathbf{x}_i^c), \tag{12}$$

where $EDR(\mathbf{x}_i^c) = |\mathbf{F}^s(\mathbf{x}_i^c) - \mathbf{F}^s(\mathbf{x}_i^c + \omega))|_2$ and ω is random small scalar. $d^{\text{CD}}(\cdot,\cdot)$ denotes uni-directional Chamfer distance. Eq.11 forces that the predicted skinned points ($\mathbf{x}_i^{pp}$) and corresponding normals ($\mathbf{n}_i^{pp}$) match the input posed points ($\mathbf{x}_i$) and their normals ($\mathbf{n}_i$). The regularisation term $E_{\text{reg}}^{\text{pose}}$ contains an L2 regulariser on the deformation field and neural surface feature as well as EDR term [64] which enforces spatial smoothness on the feature space.

4.4. Inference and Surface Extraction.

At the inference time, we predict the pose-dependent deformation for vertices $\mathbf{V}^c$ of our base fusion shape $\mathcal{B}^s$, and apply LBS [30] with given desired pose to obtain its location $\mathbf{V}^{pp}$ in the pose space. Because of the continuity of NSF, the fusion shape $\mathcal{B}^s$ here can be discretized with arbitrary resolution and topology. We use the original edge connectivity on fusion shape $\mathcal{B}^s$ and posed vertices $\mathbf{V}^{pp}$ to obtain the posed mesh, which ensures the coherency over different poses. Specifically for reconstruction task, where the partial point cloud is available, we freeze the deformation function $f^{\text{pose}}(\cdot)$ and fine-tune the neural surface feature via minimizing the single-directional Chamfer distance between the input partial shape and our reconstructed mesh together with the Laplacian smoothness loss [45] of the reconstructed mesh. Our NSF guarantees the coherent direct mesh output at arbitrary resolution without performing expensive marching cubes as in [7, 8, 11, 13, 43, 67, 69] or Poisson reconstruction [34, 39, 41, 81]

5. Experiments

Datasets. We evaluate the results of our method qualitatively and quantitatively on single-view point cloud obtained from monocular depth sequences. We rendered the depth sequences from the BuFF [80, 56] dataset and the CAPE [40, 56] dataset using Kinect camera parameters, same as our baselines [6, 13] and unproject monocular depth to use as our input along with the SMPL poses. For real data, we use Kinect depth sequences provided in DSFN [6]. We experiment with loose garments like skirts from the Resynth [41, 39] dataset.

Metrics. To evaluate the error of our method we will rely on Chamfer distance (in cm), the normal correctness, and the IoU between the ground-truth mesh and the reconstructions of our body model. The formulation of our metrics can be found in supp. material. These evaluation metrics are also applied to our baselines [6, 13].

Baselines. The work closest to ours is PINA [13] as they have the same problem setting. DSFN [6] is another baseline that uses neural network to learn SMPL-based 3D avatars from monocular RGB-D video. Since the code of PINA and DSFN is both not released, we train our model using the same data and compare with the pre-computed results provided by authors. We also compare with POP [41], MetaAvatar [69], and NPMs [49] on CAPE [40, 56] dataset. Here, we modify the Chamfer distance in POP [41] to unidirectional, allowing it accept single-view point cloud as input. Apart from these recent works, we also deploy a simple yet intuitive baseline: posing the naked SMPL shape and our learned fusion shape (w/o NSF). These baselines highlight the importance of learning pose-dependent deformations in NSF.

Table 1: We evaluate our method on the task of reconstructing 3D shape from monocular depth point clouds on BuFF [80], CAPE [40], and synthesized ReSynth [41] data. Our method performs better than existing methods both quantitatively and qualitatively.

Method	BuFF Data [80]			CAPE Data [40]			Resynth Data [41]		
	CD (cm) ↓	NC ↑	IoU ↑	CD (cm) ↓	NC ↑	IoU ↑	CD (cm) ↓	NC ↑	IoU ↑
DSFN [6]	1.56	0.916	0.832	-	-	-	-	-	-
PINA [13]	1.10	0.927	0.879	**0.62**	0.906	**0.941**	-	-	-
Ours, w/o deformation	0.97	0.922	0.851	0.86	0.929	0.869	1.14	0.915	0.846
Ours, complete	**0.69**	**0.930**	**0.895**	0.65	**0.940**	0.911	**0.92**	**0.917**	**0.887**

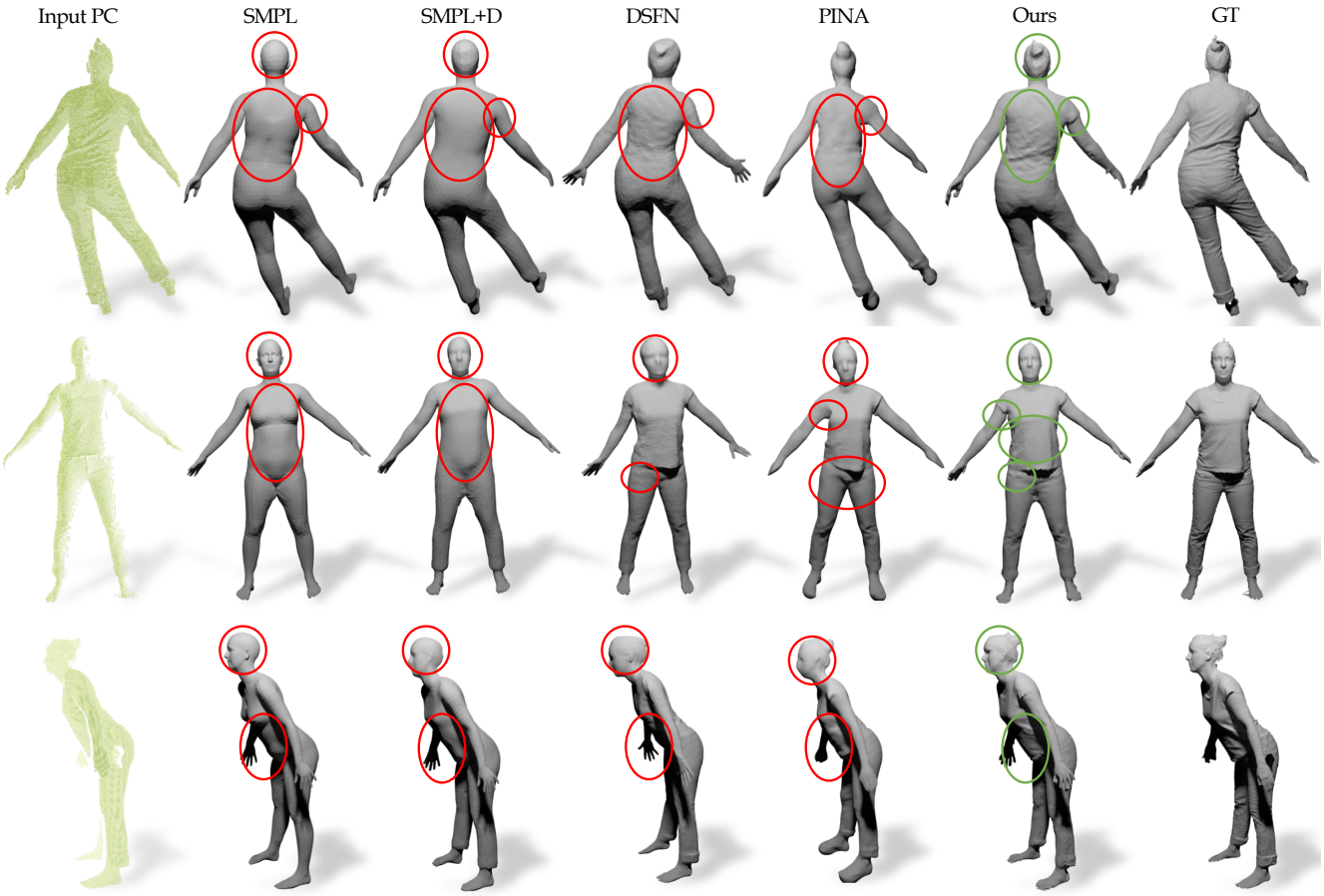

Figure 4: Partial point cloud reconstruction on BuFF [80]: We first compare with fitting SMPL and SMPL+D models to our partial point clouds and then compare against more contemporary baselines DSFN [6] and PINA [13]. Our method reconstructs more detailed avatars.

5.1. Reconstruction Comparison with Baselines.

We test our method on the task of partial point cloud reconstruction. Given a sequence of a monocular point cloud, our goal is to recover a full clothed body model. Results are reported in Tab. 1 and Fig. 4, 5. The results for each individual outfit of our method can be found in supp. material. While the competing approaches [6, 13] train a neural network per-subject, our method which is trained across multiple subjects, produces more reliable reconstructions with far less computational resources. Most essentially, our approach can reconstruct a sequence of coherent meshes at arbitrary resolution without retraining as in Fig. 1, which is not achievable by any of our baselines.

5.2. Efficiency of Neural Surface Field.

For this experiment we train 3 variants of our method with same neural networks and data but using three different feature representations, *i.e.* volume [9], tri-plane [64] and neural surface features. We report our results in Tab. 2. Our key idea to learn a deformation field on a neural surface is powerful and we can achieve better quality results with

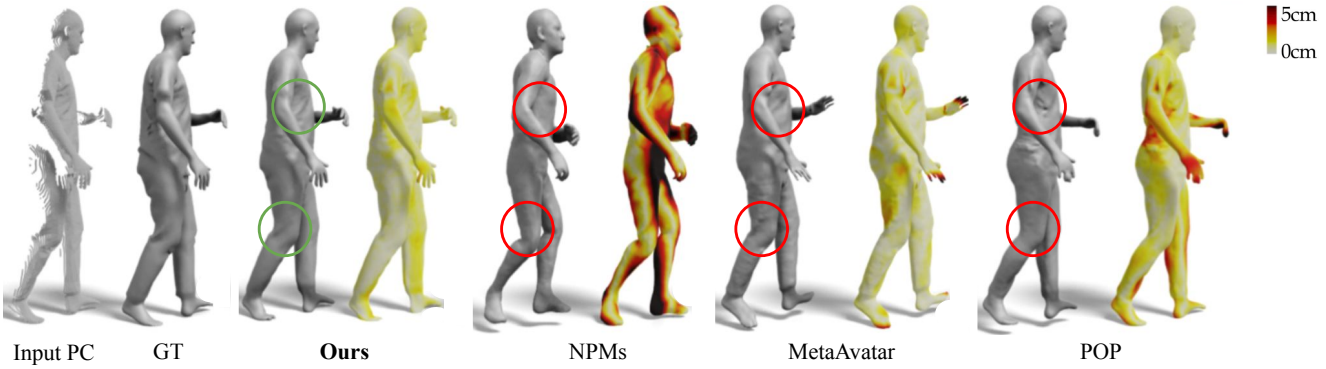

Figure 5: Partial point cloud reconstruction on CAPE [40, 56]: We compare with baselines NPMs [49], MetaAvatar [69], POP [41] and visualize the reconstruction error on the surface. Our method achieves better reconstruction quality on this dataset.

Table 2: We compare our neural surface feature learning with existing volumetric [9], and tri-plane [64] feature representation. We show that we require significantly lower learnable parameters and produce better results.

Method	BuFF Data [80] - Subject 00032			
	# Features	CD ↓	NC ↑	IoU ↑
Volume	262,144	0.77	0.925	0.884
Triplane	49,152	0.74	0.924	0.885
Ours, NSF	**6,890**	**0.66**	**0.928**	**0.899**

Table 3: Our feature decoupling allows us to use our pre-trained network and quickly learn new subject specific features with little data and time. We show that in 10 mins, and by just using 10 frames (A) from a sequence, our model achieves similar performance as training on all the frames in 10 hrs (B).

Operation	BuFF Data [80] - Subject 00114				
	# Frames	Time	CD ↓	NC ↑	IoU ↑
(A) Train	126	$\sim 600'$	0.80	0.929	0.881
(B) Fine tune	10	$\sim 10'$	0.87	0.907	0.870

$10-100$x less learnable features compared to volumetric and tri-plane features.

Moreover, by avoiding per-frame surface extraction, NSF achieves from $\sim$ 40x to $\sim$ 180x faster compared to competitors at inference time. Please refer to supp. mat. for more detail.

5.3. Importance of Feature Fecoupling: Learning a New Avatar with 10 images in under 10 mins.

Our baselines [6, 13] require training a new neural network for each subject. This is both computationally and data expensive. Our method decouples generalizable neural networks and subject-specific features, and hence we can quickly learn new subject-specific features with small amounts of data, (*i.e.* 10 depth images) in a short time ($<$ 10mins). Training a full neural network on the other hand requires several hours (see Tab. 3). We use 3 subjects from BUFF dataset for training and use 10 random frames from the 4^{th} unseen subject for learning the body model. Our qualitative results in Fig. 6 show that our decoupling allows us to learn models of new subjects easily with small amounts of data. Competing baselines [6, 13] lack such capabilities, although their code is not available for fair comparison. In our supplementary material, we also show that the generalizable decoder achieves superior performance compared to subject-specific decoder training.

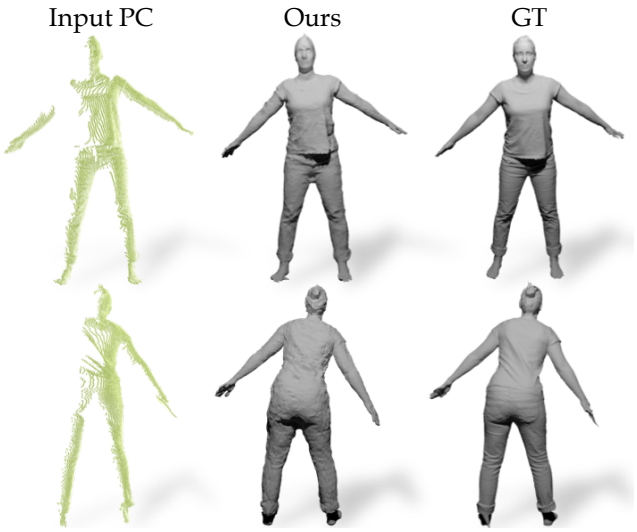

Figure 6: Point cloud reconstruction results: We learn the body model of a new subject given 10 frames in under 10 mins.

5.4. Animating Learnt Avatars.

Our method can be efficiently used to manipulate the learnt model to unseen poses. This can be done by providing the desired input pose parameters to our method. We use our model trained on BUFF [80] and animate it with poses from AIST dataset [68]. Fig. 7 shows our learnt avatars in different poses. See supp. video and pdf for more examples.

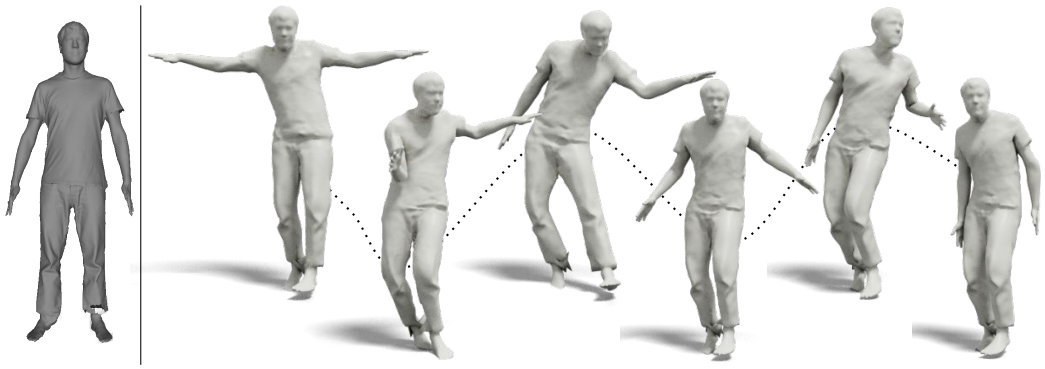

Figure 7: Since our method learns a body model of the subject, we can use this model for re-animation. We show a reference scan of a person (left) and re-posed avatars of the subject (right). Note that NSF can directly output coherent animated meshes at arbitrary desired resolution (as in Fig. 1) without retraining, which is more flexible compared to state-of-the-art works.

5.5. Results on Real Data.

In this experiment we test the generalization capability of our method on real data [6]. Fig. 8 demonstrates one example on real dataset from DSFN [6]. Both the methods are trained using same data and we our method clearly outperforms the baseline. Please supp. mat. for more examples.

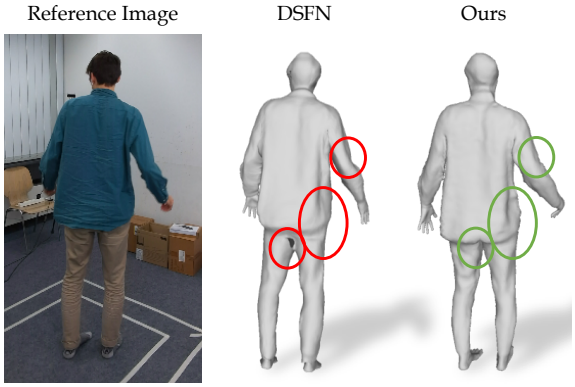

Figure 8: Generalization to real data: We show qualitative comparison with DSFN [6] on their dataset captured using a Kinect. Our model generates more details and less artefacts. Note that the reference RGB image is not used in training.

5.6. Learning Textured Avatars.

We build our fusion shape by fusing multiple monocular point clouds and our canonicalization procedure ensures that we have explicit correspondence between the input posed space and the fusion shape. This allow us to directly lift the texture from the input point cloud onto the canonical shape and we obtain a textured body model of a person. Our baselines [6, 13] have not shown such capabilities. Fig. 9 shows examples of our learnt textured avatars.

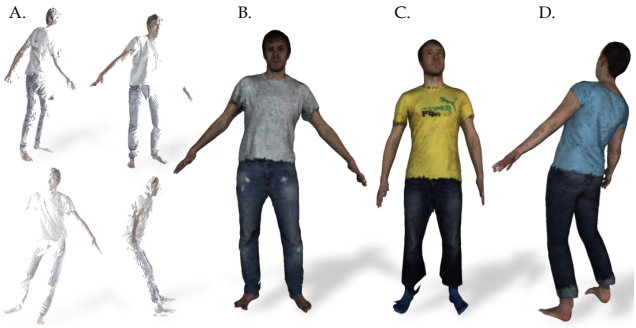

Figure 9: We can learn textured 3D avatars of people from input partial point clouds. We show sample partial inputs (A) and corresponding learnt model (B) We show more avatars in C,D.

6. Conclusion

We introduced *Neural Surface Fields* (NSF): efficient, fine-grained manifold-based continuous fields for modeling articulated clothed humans. NSF is capable of reconstructing meshes with arbitrary resolution without retraining while maintaining mesh coherency. NSF eliminates the expensive per-frame surface extraction, is about 40 to 180 times faster at inference time compared to baselines. NSF is compact and preserve the 3D structure of the underlying manifold. NSF also enables applications like texture transfer and fine-tuning to adapt to a new subject. Our evaluation on rendered and captured data demonstrate the efficiency and the power of our proposed NSF . We believe NSF can lead to both real-world applications and useful tools for the 3D vision community. The code as well as models are available at https://yuxuan-xue.com/nsf for research purposes.

Acknowledgements We appreciate Y. Xiu, G. Tiwari, H. Feng, Y. Feng for their feedbacks to improve the work. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (EmmyNoether Programme, project: Real Virtual Humans) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors

thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Y.Xue. G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. R. Marin has been supported by Alexander von Humboldt Foundation Research Fellowship and partially from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101109330.

References

[1] 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. 2, 3

[2] 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, aug 2020. 2, 3, 5

[3] 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. 2, 3

[4] 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, 3

[5] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 3

[6] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction, 2021. 2, 3, 6, 7, 8, 9

[7] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *CVPR*, 2022. 3, 6

[8] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4, 6

[9] 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. 3, 5, 7, 8

[10] 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. 5

[11] 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. 3, 6

[12] Nina Döllinger, Erik Wolf, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. Are embodied avatars harmful to our self-experience? the impact of virtual embodiment on body awareness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023. 2

[13] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. *arXiv*, 2022. 2, 3, 6, 7, 8, 9

[14] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. *CoRR*, abs/2309.06441, 2023. 3

[15] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22, 2022. 3

[16] Marie Luisa Fiedler, Erik Wolf, Nina Döllinger, Mario Botsch, Marc Erich Latoschik, and Carolin Wienrich. Embodiment and personalization for self-identification with virtual humans. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 799–800. IEEE, 2023. 2

[17] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*, 2020. 4

[18] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 2

[19] Marc Haberman, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. In *Symposium on Computer Animation(SCA)*, August 2023. 2

[20] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11046–11056, 2021. 2

[21] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37:185:1–185:15, Nov. 2018. Two first authors contributed equally. 2

[22] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *Pattern Recognition*, Lecture Notes in Computer Science, 13485, pages 281–299, Cham, Sept. 2022. Springer. 2

[23] Yangyi Huang, Yuliang Xiu, Hongwei Yi, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided

Reconstruction of Lifelike Clothed Humans. *arXiv preprint: 2308.08545*, 2023. 2

[24] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3090–3099, 2020. 2

[25] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 402–418. Springer, 2022. 3

[26] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 3

[27] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2

[28] Hyomin Kim, Hyeonseo Nam, Jungeon Kim, Jaesik Park, and Seungyong Lee. Laplacianfusion: Detailed 3d clothed-human body reconstruction. In *Proceedings of the ACM (SIGGRAPH Asia)*, 2022. 2, 3

[29] Lukas Koestler, Daniel Grittner, Michael Moeller, Daniel Cremers, and Zorah Lähner. Intrinsic neural fields: Learning functions on manifolds. In *European Conference on Computer Vision*, pages 622–639. Springer, 2022. 3

[30] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000. 6

[31] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. 2022. 4

[32] Zhe Li, Tao Yu, Zerong Zheng, Kaiwen Guo, and Yebin Liu. Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2021. 2

[33] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! Text to Animatable Digital Avatars. *arXiv preprint: 2308.10899*, 2023. 2

[34] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *ECCV*, 2022. 2, 3, 4, 6

[35] Zhen Liu, Yao Feng, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. In *International Conference on Learning Representations*, 2023. 2

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3, 5

[37] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH*, pages 163–169. ACM, 1987. 2, 5

[38] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J Black. Scale: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16082–16093, 2021. 3

[39] 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, 3, 6

[40] 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 *Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 6, 7, 8

[41] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021. 2, 3, 6, 7, 8

[42] 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

[43] 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. 3, 6

[44] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3

[45] Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. Laplacian mesh optimization. In *Proceedings of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*, GRAPHITE '06, page 381–389, New York, NY, USA, 2006. Association for Computing Machinery. 6

[46] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[47] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 2, 4

[48] Phong Nguyen, Nikolaos Sarafianos, Christoph Lassner, Janne Heikkila, and Tony Tung. Free-viewpoint rgb-d human performance capture and rendering. In *ECCV*, 2022. 2

[49] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *ICCV*, 2021. 2, 6, 8

[50] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. Spams: Structured implicit parametric models. In *CVPR*, 2022. 2

[51] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation, 2019. 3, 4

[52] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: 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. 3

[53] 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. 2, 3

[54] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 3

[55] Ilya A. Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. 2

[56] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. Two first authors contributed equally. 2, 3, 6, 8

[57] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 34(4):120:1–120:14, Aug. 2015. 3

[58] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *arXiv preprint arXiv:2306.07280*, 2023. 2

[59] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 3

[60] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 2, 5

[61] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 2, 5

[62] 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

[63] 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. 3

[64] J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion, 2022. 6, 7, 8

[65] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, October 2022. 5

[66] 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. 3

[67] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *International Conference on Computer Vision (ICCV)*, October 2021. 2, 3, 6

[68] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, Delft, Netherlands, Nov. 2019. 8

[69] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *Advances in Neural Information Processing Systems*, 2021. 2, 3, 6, 8

[70] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022. 3

[71] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, October 2022. 2

[72] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[73] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 2022. 3

[74] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Obtained from Normals. *arXiv:2212.07422*, 2022. 2

[75] 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

[76] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[77] Yuxuan Xue, Haolong Li, Stefan Leutenegger, and Joerg Stueckler. Event-based non-rigid reconstruction from contours. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 78. BMVA Press, 2022. 2

[78] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. 2

[79] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, June 2018. 2

[80] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3, 6, 7, 8

[81] Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. Closet: Modeling clothed humans on continuous surface with explicit template decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3, 6

[82] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7743–7753, June 2022. 3

[83] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021. 2

[84] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object correspondence to hand for motion refinement. In *European Conference on Computer Vision (ECCV)*. Springer, October 2022. 3

[85] Pierre Zins, Yuanlu Xu, Edmond Boyer, Stefanie Wuhrer, and Tony Tung. Data-driven 3d reconstruction of dressed humans from sparse views. In *2021 International Conference on 3D Vision (3DV)*, pages 494–504. IEEE, 2021. 3