NSF: Neural Surface Fields for Human Modeling from Monocular Depth

Yuxuan Xue\textsuperscript{1,2,*}, Bharat Lal Bhatnagar\textsuperscript{1,2,3,4,*}, Riccardo Marin\textsuperscript{1,2}, Nikolaos Sarafianos\textsuperscript{4}, Yuanlu Xu\textsuperscript{4},
Gerard Pons-Moll\textsuperscript{1,2,3,†}, Tony Tung\textsuperscript{4},
\textsuperscript{1} Tübingen AI Center \textsuperscript{2} University of Tübingen
\textsuperscript{3} Max Planck Institute for Informatics \textsuperscript{4} Meta Reality Labs Research
\textsuperscript{†} Project Lead

\{yuxuan.xue, riccardo.marin, gerard.pons-moll\}@uni-tuebingen.de
\{bharatbhatnagar, nsarafianos, yuanluxu, tony.tung\}@meta.com

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: \url{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 [50, 51, 63, 64, 66, 40] have opened new opportunities for capturing and digitising human appearance. Building digital avatars has found applications in behavioural [12, 14] and medical [71, 55] studies as well. 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, 27, 31, 33, 52, 58]. 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, 22, 67, 68]. 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 [28] and its successors [1, 4, 45, 65, 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 [27, 31, 33, 70]. Recently, point based neural implicit representations [8, 13, 52, 58, 60, 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 [29, 21] 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, 22] 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 ~ 40x and ~ 180x faster compared to point-based works [27, 31, 33, 70] using Poisson reconstruction and implicit-based works [8, 13, 52, 58, 60] 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, 22, 26, 32, 47], showing better performance than competitors, even when some of them requires subject-specific training [6, 13, 33, 41, 42, 60]. 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 [50, 51, 64, 63, 18, 73, 17] can reconstruct humans from a single RGB image but are not as accurate. Methods such as KinectFusion [39] and DynamicFusion [38] fuse depth measurements over time to create a complete and accurate model. While these are general and not restricted to humans, BodyFusion [67] and DoubleFusion [68] 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 unavail-
able. 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) [36] 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 [16, 19, 46, 53, 61, 72]. 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, 35, 58, 60, 75, 57]. 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 [28] 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 [49], face [7], soft-tissues [48] and clothing [44, 47, 48, 69, 4]. Many works model deformations [6, 22, 31, 33, 30, 69, 70, 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. [27] 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 [27, 31, 33] 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 [62]:

\[ f_\phi : \mathbb{R}^m \rightarrow \mathbb{R}^n, \]

where \( \phi \) are the learnable parameters. Neural field defined in Euclidean space \( \mathbb{R}^3 \) has been widely-used to represent various geometries like distance [43], occupancy [34], and radiance [36] functions, correspondences [2], contacts [20, 5, 74], 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 \( \mathcal{S}^2 \):

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

We call this representation *Neural Surface Fields* (NSF). Recent work [23] 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.

**Input.** Let \( s = \{1, \ldots, N\} \) be the set of subjects. For each subject, our method takes as input a sequence of monocular depth point clouds, \( \mathcal{X}^s = \{X^s_1, \ldots, X^s_T\} \). Each \( X^s_t \) 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 \( \theta^s = \{ \theta_1^s, ..., \theta_T^s \} \).

**Output.** Our goal is to learn subject-specific body models, \( M = \{ M^1, ..., M^N \} \). 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 \( \theta \), 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 \( B^s \) (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 \( \mathbf{X}_i^t \) input point cloud to a neutral pose. The corresponding canonical points \( \mathbf{X}_i^c \) for input points can be found using iterative root finding \([8, 25]\):

\[
\arg \min_{\mathbf{X}_i^c} \sum_{t=1}^{T} \left( \sum_{i=1}^{K} w(\mathbf{X}_i^c) \cdot \mathbf{T}_i(\theta_t) \right) \left( \mathbf{X}_i^c - \mathbf{X}_i^t \right) \tag{3}
\]

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 \([27]\) to avoid ambiguous solutions. We unpose all input observation \( \mathbf{X} = \{ \mathbf{X}_i \}_{t=1}^{T} \) into canonical partial shapes \( \mathbf{X}^c = \{ \mathbf{X}_i^c \}_{t=1}^{T} \).

**Implicit Fusion Shape.** Since the inverse skinning does not account for pose-dependent deformations operates at a human level, the point cloud \( \mathbf{X}_i^c \) 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 \( \{ \mathbf{X}_i^c \}_{t=1}^{T} \) by learning an implicit surface in the canonical space. Concretely, we represent \( \mathcal{B}^s \) as an implicit SDF in \([43]\), composed by a neural network \( f_{\text{shape}}(\cdot|\phi_{\text{shape}}) \) parameterised by parameters \( \phi_{\text{shape}} \), 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 \( \phi_{\text{shape}} \), are optimised with the self-supervised objective \([15]\) below:

\[
E_{\text{shape}}(\phi_{\text{shape}}, \mathcal{H}) = E_{\text{geo}} + \lambda_1 E_{\text{eik}} \tag{4}
\]

\[
E_{\text{geo}}(\phi_{\text{shape}}, \mathcal{H}) = \sum_{s=1}^{N} \sum_{t=1}^{T} \sum_{i=1}^{L_{s,t}} \left( f_{\text{shape}}(\mathbf{x}_i^c, \mathbf{h}^s|\phi_{\text{shape}}) \right) + \lambda_3 |\nabla_{\mathbf{x}} f_{\text{shape}}(\mathbf{x}_i^c, \mathbf{h}^s|\phi_{\text{shape}}) - \mathbf{n}_i^c|_2^2 \tag{5}
\]

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. ??, and \( \nabla_{\mathbf{x}} \) denotes the spatial derivative. We compute the normal \( \mathbf{n}_i \) on the point cloud using \([39]\). The term \( E_{\text{eik}}(\cdot) \) [15] enforces that the SDF prediction on the canonical surface should be zero and its derivative, i.e. normal direction, should match the canonicalised normal:

\[
E_{\text{eik}}(\phi_{\text{shape}}, \mathcal{H}) = \sum_{s=1}^{N} \sum_{t=1}^{T} \sum_{i=1}^{L_{s,t}} \left( |\nabla_{\mathbf{x}} f_{\text{shape}}(\mathbf{x}_i^c, \mathbf{h}^s|\phi_{\text{shape}})|_2^2 - 1 \right)^2 \tag{6}
\]
Insights. Our objective $E^{\text{shape}}(\phi^{\text{shape}}, \mathcal{H})$ 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 $h^s$, 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 $B^s$:

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

where points on the surface $S^2$ 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 $F = \{F^s\}_{s=1}^N$, 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:

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

where $F^s(x^c)$ denotes the latent feature queried at point $x^c$ for subject $s$ and $\theta$ 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 $B^s \subset \mathbb{R}^3$ for each subject. This requires addressing two key challenges: how to learn features $F^s(\cdot)$ 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, 50, 51] 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 $B^s$ by Marching Cubes [29] to extract an explicit surface. Moreover, if the garments can be represented by SMPL [28] 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 $F^s(x^c)$ at arbitrary surface point $x^c \in B^s$ 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 $128^3 \sim 2$mil. 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 $x^c$ to lie on the surface $B^s$ as the NSF is not even defined outside in $\mathbb{R}^3$. This is challenging because the canonical point $x^c$ obtained by canonicalising the input observation $x$ (Eq. ??) 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 $B^s$ [10, 56]. 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

---

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.
the canonical surface point $\mathbf{x}^c$ corresponding to $\mathbf{x}$.

$$
\mathbf{x}^c = \mathbf{x} + f_{\text{shape}}(\mathbf{x}, \mathbf{h}^s | \phi_{\text{shape}}) \nabla_x f_{\text{shape}}(\mathbf{x}, \mathbf{h}^s | \phi_{\text{shape}}).
$$

(9)

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

4.3. Self-supervised Cycle Consistency

Reposing via Skinning. Once we obtain the pose-dependent deformation on the fusion shape in 8, we use standard linear blend skinning [24], to reposition:

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

(10)

where $\mathbf{X}^p = \{ \mathbf{x}^i \}_{i=1}^{L_1}$ is the NSF predicted pose-dependent canonical points and $\mathbf{X}^{pp} = \{ \mathbf{x}^{pp}_i \}_{i=1}^{L_1}$ is the reposed pose-dependent points. Note that $\mathbf{X}^{pp}$ can be considered as the reconstruction of input observation $\mathbf{X}_i$.

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 $\mathbf{X}_i$. This can be formulated as the following self-supervised objective:

$$
E_{\text{pose}}(\phi_{\text{pose}}, \mathcal{F}) = \sum_{s=1}^{N} \sum_{t=1}^{T_s} \sum_{i=1}^{L_s} \left( | \mathbf{x}_i - \mathbf{x}^{pp}_i |_2 + | \mathbf{n}_i - \mathbf{n}^{pp}_i |_2 ight.
$$

$$
+ d^{CD}(\mathbf{x}_i, \mathbf{x}^{pp}_i) + E_{\text{reg}}^{\text{pose}},
$$

(11)

$$
E_{\text{reg}}^{\text{pose}} = | \mathbf{x}_i - \mathbf{x}^{pp}_i |_2 + | \mathbf{F}^s(\mathbf{x}^c_i) |_2 + EDR(\mathbf{x}^c_i),
$$

(12)

where $EDR(\mathbf{x}^c_i) = | \mathbf{F}^s(\mathbf{x}^c_i) - \mathbf{F}^s(\mathbf{x}^c_i + \omega) |_2$ and $\omega$ is random small scalar. $d^{CD}(\cdot, \cdot)$ denotes uni-directional Chamfer distance. Eq.11 forces that the predicted skinned points ($\mathbf{x}^{pp}_i$) and corresponding normals ($\mathbf{n}^{pp}_i$) 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 [54] 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 $\mathbf{B}^c$, and apply LBS [24] with given desired pose to obtain its location $\mathbf{V}^{pp}$ in the pose space. Because of the continuity of NSF, the fusion shape $\mathbf{B}^c$ here can be discretized with arbitrary resolution and topology. We use the original edge connectivity on fusion shape $\mathbf{B}^c$ 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 [37] of the re-constructed mesh. Our NSF guarantees the coherent direct mesh output at arbitrary resolution without performing expensive marching cubes as in [7, 8, 11, 13, 35, 58, 60] or Poisson reconstruction [27, 31, 33, 70].

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 [69, 47] dataset and the CAPE [32, 47] 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 [33, 31] 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 [33], MetaAvatar [60], and NPMs [41] on CAPE [32, 47] dataset. Here, we modify the Chamfer distance in POP [33] to uni-directional, 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.

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
Table 1: We evaluate our method on the task of reconstructing 3D shape from monocular depth point clouds on BuFF [69], CAPE [32], and synthesized ReSynth [33] data. Our method performs better than existing methods both quantitatively and qualitatively.

<table>
<thead>
<tr>
<th>Method</th>
<th>BuFF Data [69]</th>
<th>CAPE Data [32]</th>
<th>Resynth Data [33]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CD (cm) ↓</td>
<td>NC ↑</td>
<td>IoU ↑</td>
</tr>
<tr>
<td>DSFN [6]</td>
<td>1.56</td>
<td>0.916</td>
<td>0.832</td>
</tr>
<tr>
<td>PINA [13]</td>
<td>1.10</td>
<td>0.927</td>
<td>0.879</td>
</tr>
<tr>
<td>Ours, w/o deformation</td>
<td>0.97</td>
<td>0.922</td>
<td>0.851</td>
</tr>
<tr>
<td>Ours, complete</td>
<td><strong>0.69</strong></td>
<td><strong>0.930</strong></td>
<td><strong>0.895</strong></td>
</tr>
</tbody>
</table>

Figure 4: Partial point cloud reconstruction on BuFF [69]: 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.

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 [54] 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 $10 - 100x$ less learnable features compared to volumetric and tri-plane features.

Moreover, by avoiding per-frame surface extraction,
Figure 5: Partial point cloud reconstruction on CAPE [32, 47]: We compare with baselines NPMs [41], MetaAvatar [60], POP [33] 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 [54] feature representation. We show that we require significantly lower learnable parameters and produce better results.

<table>
<thead>
<tr>
<th>Method</th>
<th>Buff Data [69] - Subject 00032</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Features</td>
</tr>
<tr>
<td>Volume</td>
<td>262,144</td>
</tr>
<tr>
<td>Triplane</td>
<td>49,152</td>
</tr>
<tr>
<td>Ours, NSF</td>
<td><strong>6,890</strong></td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th>Operation</th>
<th>Buff Data [69] - Subject 000114</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Frames</td>
</tr>
<tr>
<td>(A) Train</td>
<td>126</td>
</tr>
<tr>
<td>(B) Fine tune</td>
<td>10</td>
</tr>
</tbody>
</table>

NSF achieves from ~ 40x to ~ 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 4th 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.

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 [69] and animate it with poses from AIST dataset [59]. Fig. 7 shows our learnt avatars in different poses. See supp. video and pdf for more examples.

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
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.

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.


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.

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


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


[54] 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


