LEARNING LAYERED NEURAL IMPLICIT MODEL FOR 3D AVATAR CLOTH REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modeling 3D clothed avatars is a popular topic in the computer graphics and vision area. Due to the complicated nature of realistic garments, the most concerned issue is how to represent 3D cloth shapes efficiently and effectively. A desirable cloth model is expected to preserve high-quality geometric details while establishing essential correspondences between clothes and animation-ready templates. However, by far there is no such a 3D cloth representation that can simultaneously satisfy these two requirements. In this work, we thus formulate a novel 3D cloth representation that integrating the neural implicit surface with a statistical body prior. Different from previous methods using explicit cloth primitives conditioned on the SMPL surface, we adopt a two-layer implicit function to capture the coarse and fine levels of cloth displacements, based on a parametric SMPL volume space. Our approach is aware of the underlying statistical minimal body shapes, and is also capable of modeling challenging clothes like skirts. To evaluate the geometric modeling capacity of our 3D cloth representation, we conduct both qualitative and quantitative experiments on raw scans, which indicate superior performance over existing 3D cloth representations. The effectiveness and flexibility of our 3D cloth representation is further validated in downstream applications, e.g. 3D virtual try-on.

1 INTRODUCTION

Dressing 3D avatars with exquisite clothes is an awesome feature in many digital applications, e.g., virtual reality, games, and fashion design. However, current 3D clothes production process heavily relies on massive labor effort as well as the expertise and aesthetics of 3D artists, as most delicate clothes nowadays we see on digital humans are manual-crafted. Therefore geometry reconstruction methods (Newcombe et al., 2015; Schönberger et al., 2016) are employed to scan and capture realistic clothes wearing on mannequins (Collet et al., 2015; Zhang et al., 2017; Pons-Moll et al., 2017; Alldieck et al., 2019; Chen et al., 2021a) as a way to improve the production efficiency and quality of 3D clothes. Nevertheless, their reconstructed raw scan results cannot be directly applied to downstream tasks such as animation, and post-scanning methods for 3D clothes extraction are still required.

There are many challenges in 3D clothes extraction from raw scan results, including the loose conditions of clothes like skirts and tuxedos, the inconsistent topology settings for one-piece and separate suits, as well as the unseen gaps between clothes and its underlying naked body. While earlier methods resort to either physics-based simulation tools (De Aguiar et al., 2010; Guan et al., 2012; Yu et al., 2019; Patel et al., 2020) or geometry processing techniques like non-rigid deformation registration (Huang et al., 2008; Li et al., 2008) to overcome these challenges, they are either restricted by specific clothes types or struggling to maintain fine clothes details such as wrinkles and folds.

Recently learning-based methods have emerged as a promising solution to address the challenges of 3D clothes extraction, which can be categorized into two major branches, namely template-based (Gundogdu et al., 2019; Patel et al., 2020) and template-free methods (Saito et al., 2021; Chen et al., 2021b; Deng et al., 2020b). To represent clothes deviation from the inner body, template-based methods define a variety of explicit geometry primitives on top of the inner body represented as SMPL (Loper et al., 2015), including displacement vectors of SMPL vertexes (Ma et al., 2020), displacement vectors of local patches (Ma et al., 2021a). Although templates provide strong priors

of clothes topologies, they also restrict the generality of template-based methods when capturing clothes details that are far from the templates. Alternatively, template-free methods adopt implicit clothes representations including occupancy fields (Mescheder et al., 2019) and signed distance fields (Fedkiw & Osher, 2002) that support flexible topology during optimization (Bhatnagar et al., 2020; Saito et al., 2021; Lin et al., 2022). From the perspective of geometric learning, these template-free methods omit dense global correspondences with generic human body models, and thus could only be over-fitted on single cases but are incapable of extracting adaptable cloth layers for different identities.

In this work, we propose a novel layered neural implicit model for 3D avatar cloth representation that combines the merits of both template-free and template-based methods, while avoiding their issues mentioned above. Specifically, our model adopts two key ideas. 1) To maintain the global correspondences with estimated SMPL shapes, similar to template-based methods our model also conditions on SMPL. But instead of conditioning on the 2D UV atlas of SMPL surface, it conditions on a 3D *uvn* space defined on SMPL surface that expands the cloth parametric space to neighboring 3D volumes. The volumetric SMPL prior is compatible with neural implicit fields, making it possible to establish global correspondences on template-free cloth representations. 2) Subsequently, to capture the high quality of clothed body scans as in template-free methods, a double-layered neural implicit function defined on the 3D uvn space is used to capture the cloth deviations from the inner body, which employs a low-frequency SIREN (Sitzmann et al., 2020) for coarse clothes structure and a high-frequency SIREN for fine clothes details.

With the help of 3D volumetric cloth space and layered neural implicit cloth representation, the proposed model not only amplifies the association between the cloth representation and the strong SMPL prior, but also enhances the grasp of high fidelity geometry details of captured cloth surfaces, especially when the clothes are far from the inner body or contain overlapping/folding structures (Chen et al., 2021b; Saito et al., 2021). After processing on a single clothed body scan, our model can be further applied to several scenarios, for instance, retargeting the clothes to different body shapes, or animating the clothes with diffused SMPL skinning weights (Lin et al., 2022). Extensive experiments are conducted to evaluate against representative baselines the geometric modeling capacity of our model in terms of the global surface reconstruction accuracy and the local detail preserving quality.

2 Related Work

Clothed avatar reconstruction requires various techniques. We mainly separate these techniques into two aspects, (i) representation of cloth for 3D avatar and (ii) neural implicit geometry learning. The following reviews illustrate recent works in detail.

2.1 3D Avatar Cloth Representation

Surface meshes are the efficient 3D representation and the predominant choice to model 3D shapes, such as the shape of common objects, cloth, and human bodys. To model cloth, previous methods either deform meshes (Bhatnagar et al., 2019; Burov et al., 2021; Ma et al., 2020; Su et al., 2022; Neophytou & Hilton, 2014; Yang et al., 2018) from an unclothed minimal body (Anguelov et al., 2005; Hirshberg et al., 2012; Loper et al., 2015; Pavlakos et al., 2019) or separate the cloth as an additional layer (Yang et al., 2018; Gundogdu et al., 2019; Liu et al., 2019; Saito et al., 2021; Patel et al., 2020) from these pre-defined templates. While these works model the cloth of 3D avatars successfully, mesh representation always suffers from the fixed topology and template registration as pre-processing. The fixed topology restricts learning general representation to various cloth categories and thus limits the generalization ability of these models and the flexibility of usage in different scenarios. Although recent works (Pan et al., 2019; Zhu et al., 2020) try to melt this limitation by adaptive template, the pre-registration between training data and proposed templates still remains challenging problems. Point clouds are another widely-used 3D representation that supports arbitrary topology and structures. Recent works (Ma et al., 2021b;a; Groueix et al., 1802) generate dense point clouds from sparse observations by patch grouping or involving features from UV maps. However, the performance of these methods is still limited by the underlying topology of the observed sparse point clouds or their template models (*e.g.* SMPL and SMPL-X (Loper et al., 2015; Pavlakos et al., 2019)).

To solve these problems, recent works attempt to involve neural implicit functions for the 3D geometry of clothes. This implicit representation does not require any pre-defined template for geometry, and thus make it flexible to reconstruct (Bozic et al., 2021; Huang et al., 2020; He et al., 2021; Saito et al., 2019; 2020; Zheng et al., 2021) and model (Saito et al., 2021; Chen et al., 2021b; Sitzmann et al., 2020; Deng et al., 2020b; Mihajlovic et al., 2021; Palafox et al., 2021) shapes of 3D human under numerous different surface topology. Despite these methods having the theoretical capacity to handle varied cloth topology and realistically structure details, it remains challenging to design learning and representation strategies to realize the full potential of this representation (*e.g.* reconstruction on details of clothes). Moreover, the generalization ability to different scenarios (*e.g.* body shapes and poses) of these methods still requires further exploration.

2.2 Neural Implicit Geometry Learning

Recently, the neural implicit function has shown the capacity to reconstruct high-quality 3D geometries (Park et al., 2019; Chen & Zhang, 2019; Mescheder et al., 2019). Although these neural implicit function based methods can model geometry with infinite resolution theoretically, the reconstructed target is still independent of the representation ability of neural networks in practice. To improve the representation and generalization ability of the neural network to 3D geometries (*e.g.* details of 3D surfaces), on the one hand, many works have introduced spatial-wise hierarchical information to model these structures. (Chabra et al., 2020; Saito et al., 2019; 2020) use the hybrid representation with sparse voxels and dense 2D grids to improve the detail of reconstruction. (Takikawa et al., 2021; Liu et al., 2020; Martel et al., 2021) introduce shape-adaptive octrees with learned prior codes to improve the quality of reconstruction. However, these methods always demand large memory and computation source to model structures in detail. Differently, some methods try to decompose the target shape by parameterized templates (Genova et al., 2019; Deng et al., 2020a; Chen et al., 2020), but they are still limited by the template function and delicate spatial blending issues to model detail structures.

On the other hand, there are some methods that begin to focus on the ability of networks on high-frequency signals to model exquisite details. (Sitzmann et al., 2020; Mildenhall et al., 2020) involve the high-frequency information by the positional encoding or sinusoidal representation. Moreover, (Yifan et al., 2021; Li & Zhang, 2021) introduce the decomposition to the target structure and use different neural networks to capture coarse and fine geometry structures, respectively. Our method follows this decomposition paradigm and explores how to model the coarse body and details of 3D cloth by different neural implicit functions. Besides, we evaluate that our representation not only can reconstruct cloth with rich details but also can be extended to applications like body shape retargeting.

3 Methodology

In this section, we first analyze the technical challenges of building a successful 3D cloth representation for scan based clothed avatar generation. Then we introduce our novel approach to address these issues, mainly from two aspects, the prior conditioning scheme and the implicit surface modeling techniques. Finally, we present the full pipeline as well as the loss functions to optimize our proposed 3D cloth representation on input scan.

3.1 Problem Formulation

As multi-view stereo (Schönberger et al., 2016) and depth-aware (Newcombe et al., 2015) systems have become more and more available to the 3D content generation purpose, modeling realistic human avatar directly from high-fidelity point clouds is a promising way to create digital twins. Among several technical components of this area, the cloth modeling problem is one of the most important feature remaining unsolved. In this work, we present a powerful 3D cloth representation which can be applied to produce plausible clothes from raw scans.

Specifically, given a clothed body scan $\mathcal{P}_{scan} \in \mathbb{R}^{3\times K}$ containing K surface points that we deem as sampled from continuous manifolds $\mathcal{X}_{scan}$, and a minimal body mesh $\mathcal{X}_{body}$ which can be easily estimated (Zhang et al., 2017; Zuo et al., 2021) via fitting the SMPL (Loper et al., 2015) shape and pose parameters (β, θ) by:

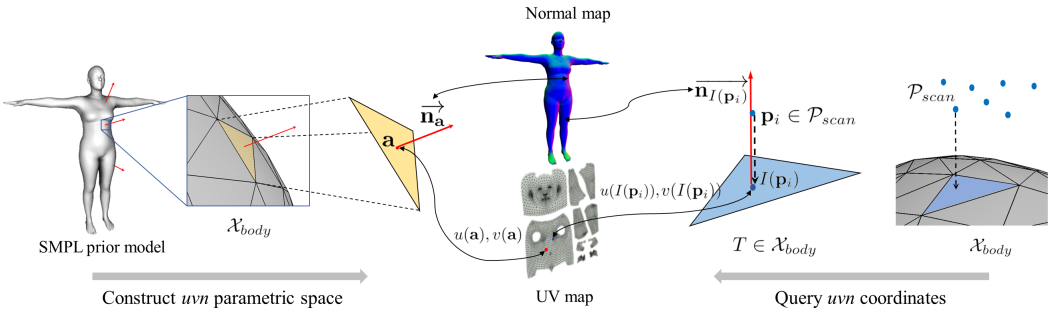

Figure 1: Illustration of our proposed volumetric prior space. From the most left to the middle, we show the the construction process of the parametric space derived from SMPL prior model. Once the uvn parametric space are established, we can query the uvn coordinates of an arbitrary points above the prior shape, following the steps from the most right to middle.

$$\begin{aligned} T(\beta,\theta) &= \bar{T} + B_S(\beta) + B_P(\theta), \\ \mathcal{X}_{body} &= W(T(\beta,\theta), J(\beta), \theta, \mathcal{W}), \end{aligned} \tag{1}$$

where $\bar{T}$, $B_S(\cdot)$, and $B_P(\cdot)$ are the mean, shape, and pose component of SMPL, $W(\cdot)$ is the blend skinning function, J is the 3D joints computed from β, and $\mathcal{W}$ is the skinning weights, the 3D cloth modeling problem is referred to represent the displacement d between the minimal and clothed body surface like: $\mathcal{X}_{cloth} = D\left(\mathcal{X}_{scan} \mid \mathcal{X}_{body}\right)$.

Despite there are quite a few cloth representation works already existed (Bhatnagar et al., 2020; Ma et al., 2020; Corona et al., 2021; Ma et al., 2021a;b), they all suffer from two challenges of this task. The first one is the constraint of statistical body prior. Due to limited mesh resolution and over-smoothed surface, conditioning the cloth layer on the SMPL model $\mathcal{X}_{body}$ may not generate sufficient correspondences to capture dense and rich cloth features, especially in complex and loose garment cases. The other challenge is about reconstructing cloth geometry from scans. Although many scanning systems have achieved high-fidelity results, most cloth representations based on sparse/discrete primitives, e.g., mesh vertices (Ma et al., 2020) or patches (Ma et al., 2021a), are still incapable of modeling high-quality surface details because they have to compromise the geometry accuracy for structure consistency.

To address the above two challenges, we propose a new 3D cloth representation, consisting of a volume-based SMPL parametric space and a neural implicit surface model. Particularly, we decompose the cloth modeling problem into two parts: (1) encoding the estimated minimal body shape into a volumetric space for cloth parameterization; (2) decoding the cloth surface by adopting a layered neural implicit function to regress the displacement fields between minimal body and cloth outfits. More technical details are introduced respectively below.

3.2 Volumetric Prior Parameterization

In order to improve the cloth modeling capacity of SMPL-based prior shapes, we introduce a volume-based approach to elevate the cloth conditioning space from 2D to 3D. By parameterizing the SMPL mesh into UV atlas, it is convenient to represent a point position on the minimal body surface with its corresponding UV coordinates. Most previous works (Ma et al., 2020; 2021a;b) bond the cloth displacements with the SMPL parametric space. These surface based representations demonstrate good performances on simple and tight clothes, but are struggling to capture dense and detailed cloth shapes, especially for loose garments far from the inner body. Therefore, we take inspiration from the cylindrical coordinate system and propose a volumetric parameterization approach to address this issue. The motivation is to increase the SMPL parametric space from the minimal body surface to a bounded volume, such that the target cloth shapes can be well captured with the 3D-aware parametric space.

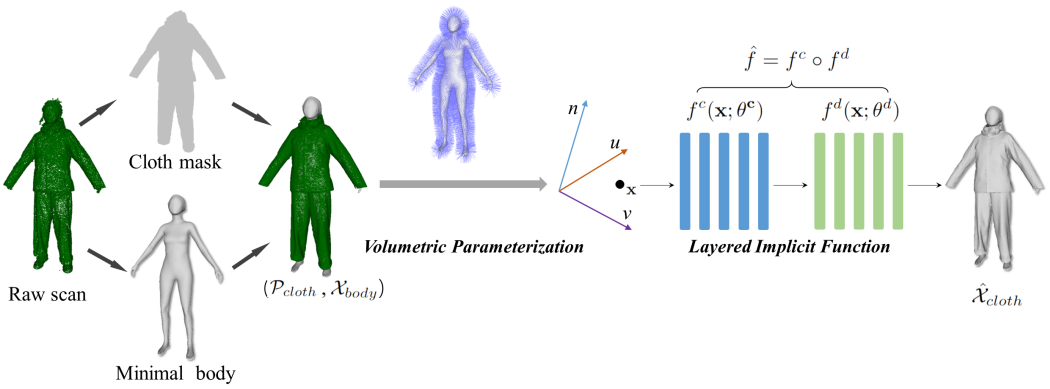

Figure 2: The full pipeline of our proposed learning framework for layered implicit cloth model. The scan data as well as minimal body estimation are from the *TightCap* (Chen et al., 2021a) dataset.

As illustrated in Fig. 1, we show the overall structure of our volumetric prior space. For each on-surface point $\mathbf{a} \in \mathbb{R}^3$ of minimal body $\mathcal{X}_{body}$, we first obtain its UV coordinates $(u(\mathbf{a}), v(\mathbf{a}))$ by mesh parameterization and barycentric interpolation, then we assign an additional scalar $n(a) \in [0, 1]$ along its normal $\overrightarrow{\mathbf{n}_\mathbf{a}}$ on mesh surface. In this way, we define a three-dimensional coordinate system (u, v, n) that can be used for representing the bounded volume above the SMPL surface. Notably, the surface normals on triangulated meshes are not naturally continuous because of the discrete facets. Therefore, we choose vertex-based normals to interpolate $\overrightarrow{\mathbf{n}_\mathbf{a}}$ by its interior barycentric coordinates on triangle. Since SMPL prior shapes are generally over-smoothed, our derived volumetric coordinates also share nice smooth property across the minimal body surface as well as in the bounded volume.

In practice, given a minimal body shape $\mathcal{X}_{body}$ and the corresponding raw scan $\mathcal{P}_{scan}$, we first build up the volumetric prior space $u \times v \times n = \{(u, v, n)\} \subseteq [0, 1]^3$ based on the parametric coordinates and normals of $\mathcal{X}_{body}$. Next we need to transfer the scan data $\mathcal{X}_{scan}$ from the world coordinates system to the local volume, finding its projected positions on $\mathcal{X}_{body}$ and computing the normal-guided displacements. An efficient way is to retrieve the closest triangle $T \in \mathcal{X}_{body}$ to every $\mathbf{p}_i \in \mathcal{P}_{scan}$, and then calculate the projection $I(\mathbf{p}_i)$ as well as its barycentric coordinates on T. By looking up the pre-defined UV map $u(\cdot), v(\cdot)$ and point normals of $\mathcal{X}_{body}$, the volumetric coordinates of $\mathbf{p}_i$ can be finally calculated as:

$$\begin{aligned} u_{\mathbf{p}_i} &= u(I(\mathbf{p}_i)) \ , \\ v_{\mathbf{p}_i} &= v(I(\mathbf{p}_i)) \ , \\ n_{\mathbf{p}_i} &= \|\mathbf{p}_i - I(\mathbf{p}_i)\| \ / \ \|\overrightarrow{\mathbf{n}_{I(\mathbf{p}_i)}}\| \ . \end{aligned} \tag{2}$$

Thanks to the consistent correspondences across SMPL meshes, the uvn coordinates are universal across different body shapes and the bounded volumes. In other words, as long as the minimal body $\mathcal{X}_{body}$ is estimated, the original point cloud $\mathcal{P}_{scan}$ defined on xyz world coordinates can be parameterized, just like the mapping from SMPL surface to UV map, to the local uvn space. Compared to the 2D prior conditioning scheme equipped with explicit cloth primitives, the 3D volumetric prior space is cooperative with implicit geometric functions like SDF, supporting flexible topology of surface regression without losing high-quality cloth details.

3.3 Layered Neural Implicit Cloth Model

Next we introduce our solution of modeling cloth shapes within the 3D volumetric space. Given a pointcloud of clothed body scan $\mathcal{P}_{scan}$, we first clean the raw scan by distilling the non-cloth area. The cloth mask can be obtained from multiple way, either in manual (Chen et al., 2021a) or in automatic (Zhang et al., 2017; Pons-Moll et al., 2017), and help to parse the cloth part as $\mathcal{P}_{cloth}$. Our basic idea is to employ the neural implicit geometry learning techniques (Chen & Zhang, 2019; Park et al., 2019; Mescheder et al., 2019), which have been proved efficient in representing object shapes

free of templates, to capture the outfit surface of cloth. Different from the previous coordinate-based implicit geometry model, our implicit cloth representation is built on the parametric space instead of the Cartesian coordinate system. In general, the cloth shape $\mathcal{X}_{cloth}$ can be represented by the iso-surface of the signed distance function (SDF) defined in the uvn space as following:

$$\mathcal{X}_{cloth} = \{\mathbf{x} \mid f(\mathbf{x}; \theta) = 0, \ \ \mathbf{x} \in u \times v \times n\} \ , \tag{3}$$

where $f : \mathbb{R}^3 \times \mathbb{R}^{|\theta|} \rightarrow \mathbb{R}$ denotes the SDF value predicting function, storing the implicit surface representation of $\mathcal{X}_{cloth}$ in a multi-layer perceptron (MLP) parameterized by θ and activated by SIREN (Sitzmann et al., 2020).

However, because of the inductive bias issue, a single neural implicit representation network is difficult to handle both smooth and detailed surface of clothes. To ease the regression difficulty of f, we propose a layered implicit representation with two SIREN blocks in our framework, trying to model the cloth shapes in a coarse-to-fine manner.

In practice, we construct a base shape network f^c with low-frequency based sine activation module θ^c, and a residual displacement network f^d with high-frequency module θ^d. A naive idea for composing the base and residual SDF estimation function is to directly compute the sum of their values, i.e., $f^c(\mathbf{x}; \theta^c) + f^d(\mathbf{x}; \theta^d)$. Unfortunately, it does not improve the shape modeling performance from the single implicit module's estimation, and even worse, the mixture of high and low frequency based network negatively affects the convergence of training and then produces unstable results. So we take inspiration from the implicit displacement fields (Yifan et al., 2021) to formulate the residual component as normal-guided displacement on the base SDF fields. Specifically, let f^c to capture the coarse surface shape $\mathcal{X}^c_{cloth}$ by storing the base SDF value:

$$\bar{\mathcal{X}}_{cloth} = \{\mathbf{x} \mid f^c(\mathbf{x}; \theta^c) = 0, \ \ \mathbf{x} \in u \times v \times n\} \ . \tag{4}$$

Meanwhile, the coarse surface normals can be calculated by estimating the gradients of f^c as:

$$\mathbf{n}(\mathbf{x}) = \frac{\nabla f^c(\mathbf{x}; \theta^c)}{\|\nabla f^c(\mathbf{x}; \theta^c)\|} \ . \tag{5}$$

Then we enforce the high-frequency component f^d to predict the detail displacement vector $\Delta\mathbf{x}$ along the normal direction $\mathbf{n}(\mathbf{x})$ as:

$$\Delta\mathbf{x} \ = \ f^d(\mathbf{x}; \theta^d) \cdot \mathbf{n}(\mathbf{x}), \tag{6}$$

and formulate the overall implicit function $\hat{f} = f^c \circ f^d$ as following:

$$\hat{f}(\mathbf{x}; \theta^c \cup \theta^d) = f^c(\mathbf{x}'; \theta^c) \ \ , \ \ \mathbf{x}' = \mathbf{x} + \Delta\mathbf{x} \ . \tag{7}$$

In this way, the detail-enhanced cloth surface can be denoted by the zero-level set of $\hat{f}$:

$$\hat{\mathcal{X}}_{cloth} = \{\mathbf{x} \mid \hat{f}(\mathbf{x}; \theta^c \cup \theta^d) = 0, \ \ \mathbf{x} \in u \times v \times n\} \ . \tag{8}$$

The network structure of $\hat{f}$ is depicted in Fig. 2. It is worth mention that, despite there is no available supervision to the individual component of implicit function f^c and f^d, the layered model $\hat{f} = f^c \circ f^d$ can be well optimized on the original point cloud data because of its self-contained structure. In ablative studies, we also demonstrate the better performances of this layered design than the baseline.

3.4 Optimization Pipeline

With the layered design of the two SIREN networks, we propose to learn the composite SDF function $\hat{f}$ directly from clothed body point cloud $\mathcal{P}_{scan}$. Before that, we assume each point $\mathbf{p} \in \mathcal{P}_{scan}$ is transformed from the Cartesian coordinates (x, y, z) to the parametric coordinates (u, v, n), and re-oriented in the volumetric space, i.e., updating the attached normal $\mathbf{n}_i$ with $\partial u, \partial v, \partial n$ derivatives.

Now we can optimize the SDF function $\hat{f}$ with the input point cloud $\mathcal{P}_{scan}$ by solving the eikonal equation (Gropp et al., 2020) . The optimizing function includes three parts. The first is the boundary condition on $(\mathbf{p}, \mathbf{n}) \in \mathcal{P}_{scan}$:

$$\mathcal{L}_{boundary}(\hat{f}) = \sum_{(\mathbf{p},\mathbf{n})\in\mathcal{P}_{scan}} |\hat{f}(\mathbf{p})| + \lambda(\langle \nabla \hat{f}(\mathbf{p}), \mathbf{n}\rangle - 1) \ , \tag{9}$$

where λ is a weight coefficient and $\langle \cdot \, , \cdot \rangle$ denotes dot product. Next is the eikonal assumption on SDF gradients:

$$\mathcal{L}_{eikonal}(\hat{f}) = \sum_{\mathbf{x}\in[0,1]^3} \big| \|\nabla \hat{f}(\mathbf{x})\| - 1 \big| \ , \tag{10}$$

$\| \cdot \|$ is the $l2$-norm and this assumption applies to the whole parametric volume $u \times v \times n$ derived from the minimal body shape. The third one is to penalize the SDF values for points that are not on the target surface:

$$\mathcal{L}_{penalize}(\hat{f}) = \sum_{\mathbf{x}\in[0,1]^3\setminus\mathcal{P}_{scan}} \exp\big(-100\hat{f}(\mathbf{x})\big) \ . \tag{11}$$

Finally, the full loss function is given with trade-off parameters $\alpha_{1,2,3}$:

$$\mathcal{L}_{full}(\hat{f}) = \alpha_1\mathcal{L}_{boundary}(\hat{f}) + \alpha_2\mathcal{L}_{eikonal}(\hat{f}) + \alpha_3\mathcal{L}_{penalize}(\hat{f}) \ . \tag{12}$$

4 EXPERIMENT

We present the evaluation and application results of our method in this section. To evaluate the cloth modeling capacity of our method, we compare with other 3D cloth representations, both in terms of the global geometry recovering accuracy and the local detail preserving quality. We also conduct ablative studies to explore the effectiveness of each component in our framework. Finally, we demonstrate some extending applications based on our proposed 3D cloth representation and the other essential techniques.

Implementation details. In practice, we use cotangent barycentric weights to interpolate the face normals of the SMPL (Loper et al., 2015) mesh to vertex based normals. The normal vector length is empirically set as 3-unit to cover most cloth deviations in our dataset. Both the low- and high-frequency SIREN modules have four layers and 256 channels in each. To avoid the detail displacement module generating large artifacts, an 0.1 scale tanh activation layer is added at the last layer of f^d. We follow the frequency picking strategy of (Yifan et al., 2021) and choose the sine frequency for f^c and f^d as 15 and 60 respectively. The weight coefficient $\lambda = 0.01$, and the trade-off parameters $\alpha_1 = 10.0$, $\alpha_2 = 0.1$, $\alpha_3 = 1.0$. We train the layered networks on single scan for 200 epochs at learning rate of $1e-4$ with ADAM optimizer. For each round, the batch size is $20,000$, including $10,000$ on-surface points from scan data and $10,000$ random samplings within uvn space. After the neural implicit network training finished, we extract the zero-level surface of cloth shapes by the marching cube algorithms performed in a 1000^3 volume.

Dataset. We test our method on various public scan dataset, including the original *CAPE* (Pons-Moll et al., 2017; Ma et al., 2020) dataset, *ReSynth* (Ma et al., 2021a;b), and *TightCap* (Chen et al., 2021a) dataset.

4.1 EVALUATION

First, we evaluate the geometry reconstruction results based on different cloth representations with the original scan. Two kinds of numerical metric are calculated, including the Chamfer-L2 distance and the normal cosine distance. In the second part, we validate the model design of the layered neural implicit function in our method. By comparing the results obtained from different choices of the two components, we demonstrate the effectiveness of our framework design.

Table 1: Quantitative results of the cloth modeling accuracy of different 3D cloth representation methods, including *CAPE* (Ma et al., 2020), *SCALE* (Ma et al., 2021a), *POP* (Ma et al., 2021b), *FITE* (Lin et al., 2022), and ours. We compare the Chamfer-L2 distance and normal difference between the generated cloth shapes and the original scans, based on two public dataset, CAPE (Ma et al., 2020) and ReSynth (Ma et al., 2021a). For each column, we give the mean and max error of each approach on the dataset. The best scores are indicated in bold.

Methods	*CAPE*		*ReSynth*	
	Chamfer-L2 ↓	Normal diff. ↓	Chamfer-L2 ↓	Normal diff. ↓
	Mean / Max	*Mean / Max*	*Mean / Max*	*Mean / Max*
CAPE	1.91 / 4.21	1.17 / 1.96	2.81 / 6.18	3.76 / 6.97
SCALE	0.98 / 3.17	1.09 / 1.77	1.96 / 4.32	2.55 / 5.12
POP	0.79 / 1.68	1.05 / 1.43	1.33 / 3.75	1.84 / 4.19
FITE	**0.76** / 1.74	**0.98** / **1.24**	1.25 / 3.35	1.49 / 3.39
Ours	**0.76** / **1.69**	1.02 / 1.27	**1.17** / **3.22**	**1.41** / **3.19**

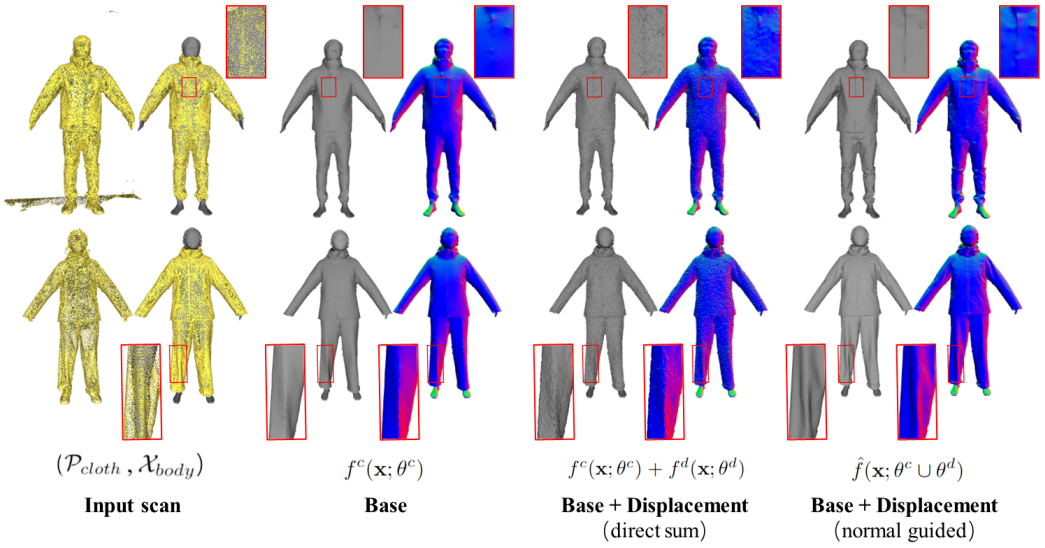

Figure 3: Ablative study results on the neural implicit model design. From left to right, we compare our layered cloth representation (the fourth) with the baseline implemented by the single SIREN module (the second) and the direct sum of low and high frequency module (the third).

Comparison with other cloth representations. We compare our cloth modeling results with other learning based approaches, including *CAPE* (Ma et al., 2020), *SCALE* (Ma et al., 2021a), *POP* (Ma et al., 2021b), and *FITE* (Lin et al., 2022). The experiment is conducted by processing single scan in the *CAPE* (Ma et al., 2020) and ReSynth (Ma et al., 2021a) dataset. The error between fitted cloth surface and original point clouds is evaluated by two metrics. The Chamfer-L2 distance represents overall shape reconstruction accuracy and the mean value of normal cosine distance reflects the local detail preserving quality. In Tab. 1, we report the mean and max error across each dataset. The numerical results indicate our method is comparable to the most recent implicit approach FITE on the CAPE dataset, and achieves the best performance on the ReSynth dataset.

Ablative study. Next we validate the component design of the layered neural implicit cloth representation. Based on the parametric volume space, we compare two baselines of using implicit functions to regress the cloth surface. A naive choice is to use a single SIREN module to optimize

the SDF prediction function f^c, however, despite the single-layer function can reconstruct the coarse cloth shapes, over-smooth problem still exists due to the inductive bias. The other baseline is to add an extra high-frequency SIREN module f^d, to enhance the detail modeling capacity of cloth representation. Unfortunately, the straightforward approach, i.e., adding the SDF values predicted by f^c and f^d, is sensitive to optimization and generally produces noisy sdf values. Following the implicit displacement fields (Yifan et al., 2021), we incorporate a normal-guided design to overcome this issue by making the high-frequency component performs in a residual manner. As shown in Fig. 3, we demonstrate two cases from the *TightCap* dataset. The rendering results show that our method can reconstruct the cloth surface smoothly as well as nice details.

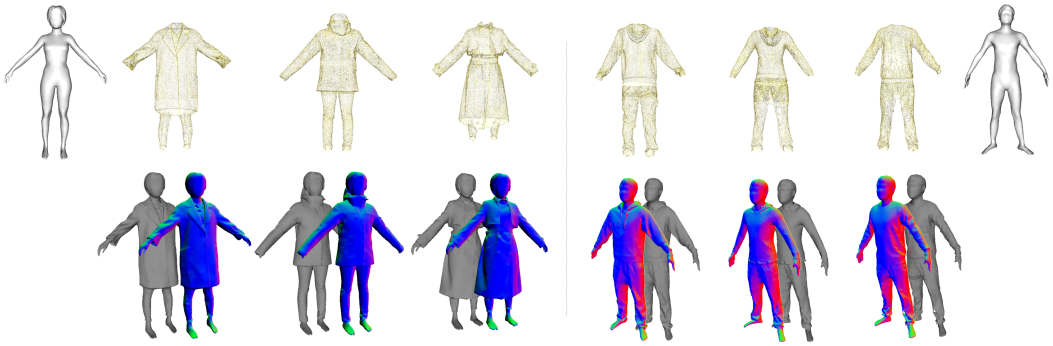

Figure 4: Cloth retargeting examples using our implicit cloth fitting model conditioned on the parametric volume space of statistical minimal body. While translating the implicit cloth shape from captured body to the target, our approach can handle simple cases like tight shirts (right) as well as challenging cases e.g. loose skirts (left).

4.2 Application

Benefited from the joint of template-based and template-free methods, our proposed neural 3D cloth representation can be applied to a variety of downstream tasks. Once an implicit cloth model is fitted on an original scan, it can be retargeted to arbitrary shapes or animated with correct pose guidance. Because the signed distance function is defined on the volume space above the estimated minimal body surface, the implicit fields can be regarded as the local attachment to the underlying body. Therefore, it is convenient to manipulate the cloth deviation by changing the global SMPL parameters. In Fig. 4, we show two example cases of 3D virtual try-on. We first build up individual neural implicit representation network for each cloth based on the original scan. After constructing the parametric volume space of a new body shape, the trained cloth model can be integrated with the new parametric volume and thus generating faithful cloth retargeting results.

5 Conclusion

Taking a step towards simultaneously capturing the high-quality geometric details of 3D clothes while establishing correspondences between clothes and the underlying human body, in this paper we propose a novel 3D cloth representation based on a layered neural implicit function. The layered neural implicit function adopts two SIREN networks to respectively model the coarse and fine cloth geometric details. Moreover, instead of defining this function on the SMPL surface as in previous methods, we defining this function on a novel volumetric SMPL space, ensuring the compatibility between the implicit clothes representation and the explicit human body template. On two challenging datasets, namely CAPE and ReSynth, the proposed 3D cloth representation not only demonstrates superior geometric modeling capacity over existing 3D cloth representations, but also shows great potential in downstream tasks like 3D cloth retargeting across different subjects.

REFERENCES

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1175–1186, 2019.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pp. 408–416. 2005.

Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *proceedings of the IEEE/CVF international conference on computer vision*, pp. 5420–5430, 2019.

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*, pp. 311–329. Springer, 2020.

Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1450–1459, 2021.

Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10754–10764, 2021.

Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *European Conference on Computer Vision*, pp. 608–625. Springer, 2020.

Xin Chen, Anqi Pang, Yang Wei, Wang Peihao, Lan Xu, and Jingyi Yu. Tightcap: 3d human shape capture with clothing tightness field. *ACM Transactions on Graphics (Presented at ACM SIGGRAPH)*, 2021a.

Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11594–11604, 2021b.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5939–5948, 2019.

Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 45–54, 2020.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015.

Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11875–11885, 2021.

Edilson De Aguiar, Leonid Sigal, Adrien Treuille, and Jessica K Hodgins. Stable spaces for real-time clothing. *ACM Transactions on Graphics (TOG)*, 29(4):1–9, 2010.

Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31–44, 2020a.

Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *European Conference on Computer Vision*, pp. 612–628. Springer, 2020b.

Stanley Osher Ronald Fedkiw and Stanley Osher. Level set methods and dynamic implicit surfaces. *Surfaces*, 44(77):685, 2002.

Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7154–7164, 2019.

Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pp. 3569–3579. 2020.

T Groueix, M Fisher, VG Kim, BC Russell, and M Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. arxiv 2018. *arXiv preprint arXiv:1802.05384*, 1802.

Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics (ToG)*, 31(4):1–10, 2012.

Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8739–8748, 2019.

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*, pp. 11046–11056, 2021.

David A Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. Coregistration: Simultaneous alignment and modeling of articulated 3d shape. In *European conference on computer vision*, pp. 242–255. Springer, 2012.

Qi-Xing Huang, Bart Adams, Martin Wicke, and Leonidas J Guibas. Non-rigid registration under isometric deformations. In *Computer Graphics Forum*, volume 27, pp. 1449–1457. Wiley Online Library, 2008.

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

Hao Li, Robert W Sumner, and Mark Pauly. Global correspondence optimization for non-rigid registration of depth scans. In *Computer graphics forum*, volume 27, pp. 1421–1430. Wiley Online Library, 2008.

Manyi Li and Hao Zhang. D2im-net: Learning detail disentangled implicit fields from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10246–10255, 2021.

Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. *arXiv preprint arXiv:2207.06955*, 2022.

Lijuan Liu, Youyi Zheng, Di Tang, Yi Yuan, Changjie Fan, and Kun Zhou. Neuroskinning: Automatic skin binding for production characters with deep graph networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020.

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.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6469–6478, 2020.

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*, pp. 16082–16093, 2021a.

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*, pp. 10974–10984, 2021b.

Julien NP Martel, David B Lindell, Connor Z Lin, Eric R Chan, Marco Monteiro, and Gordon Wetzstein. Acorn: Adaptive coordinate networks for neural scene representation. *arXiv preprint arXiv:2105.02788*, 2021.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10461–10471, 2021.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference Computer Vision*, pp. 405–421, 2020.

Alexandros Neophytou and Adrian Hilton. A layered model of human body and garment deformation. In *2014 2nd International Conference on 3D Vision*, volume 1, pp. 171–178. IEEE, 2014.

Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 343–352, 2015.

Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12695–12705, 2021.

Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9964–9973, 2019.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7365–7375, 2020.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15, 2017.

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

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 84–93, 2020.

Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2886–2897, 2021.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pp. 501–518. Springer, 2016.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11358–11367, 2021.

Shan Yang, Zherong Pan, Tanya Amert, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. Physics-inspired garment recovery from a single-view image. *ACM Transactions on Graphics (TOG)*, 37(5):1–14, 2018.

Wang Yifan, Lukas Rahmann, and Olga Sorkine-Hornung. Geometry-consistent neural shape representation with implicit displacement fields. *arXiv preprint arXiv:2106.05187*, 2021.

Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap: Single-view human performance capture with cloth simulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5504–5514, 2019.

Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4191–4200, 2017.

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.

Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *European Conference on Computer Vision*, pp. 512–530. Springer, 2020.

Xinxin Zuo, Sen Wang, Minglun Gong, and Li Cheng. Unsupervised 3d human mesh recovery from noisy point clouds. *arXiv preprint arXiv:2107.07539*, 2021.