IMPLICITPCA: Implicitly-Proxied Parametric Encoding for Collision-Aware Garment Reconstruction

Lan Chen[1,2], Jie Yang[1,†], Hongbo Fu[3], Xiaoxu Meng[4], Weikai Chen[4], Bo Yang[4], and Lin Gao[1,2]

[1]*Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
[2]*University of Chinese Academy of Sciences, Beijing, China*
[3]*City University of Hong Kong, HongKong*
[4]*Tencent America, Los Angeles, USA*

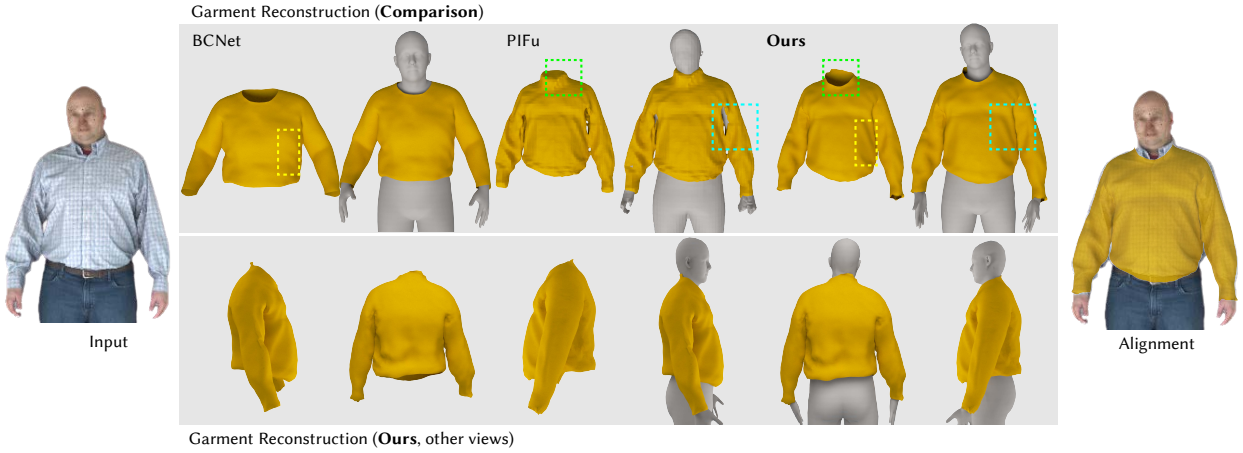


Figure 1: Taking a single-view image as input, the first row shows the reconstructed garments by BCNet [23], PIFu [40] and our IMPLICITPCA; the second row shows our inference from three other views. Our approach achieves high-fidelity details and avoids collisions with bodies.

Abstract

The emerging remote collaboration in a virtual environment calls for the need for high-fidelity 3D humans that can be quickly generated from a single image. To deal with the challenges of estimating rich clothing details and topologies, parametric models are widely used as explicit priors. Though global deformation can be achieved, this line of approach often lacks fine details from the image. Alternative approaches based on neural implicit function generate accurate details but are typically limited to closed surfaces. In addition, as human avatars are typically required to be animatable in telepresence, achieving physically correct reconstructions, e.g. collision-free, is crucial for realistic modeling but often ignored in prior works. To solve these problems, we present *ImplicitPCA*, a framework for single-view garment reconstruction that bridges the good ends of explicit and implicit representation while achieving high-fidelity and physically-plausible results. The key to our approach is a parametric SDF network that closely couples parametric encoding with implicit functions. Therefore, the parametric models can enjoy the fine details brought by implicit reconstruction while maintaining correct topology with open surfaces. To ensure physically-correct estimation, we introduce a collision-aware regression network, with fast collision detection and penalization, to capture both cloth and human signed distance fields (SDFs) from the input image. During inference, given an input image with 2D garment landmarks, an iterative routine is applied to obtain the optimal parameters by aligning the projection of the cloth mesh with the 2D landmarks and fitting the parametric implicit fields with the reconstructed collision-aware cloth SDF. The experiments on the public 3D garment dataset and in-the-wild images demonstrate that our result outperforms the prior works and

† Corresponding author: Jie Yang, yangjie01@ict.ac.cn

provides an effective tool for reconstructing detailed and topology-correct 3D garments while avoiding garment-body collisions.

Keywords: Garment Reconstruction, Implicit and Explicit Representation, Collision Aware, Parameterized Generation, Optimization

1. Introduction

Remote and distributed collaboration has become increasingly prevailing nowadays, following the global impact of Covid-19 pandemic and the surge of the metaverse. Among all the needs of digitizing our real world, clothed human models with high-fidelity and physical-plausible details are the most demanding assets in establishing a life-like virtual environment. While the undressed human is relatively easy to model due to their consistent topology, reconstructing 3D garments are much more challenging as they typically contain rich geometry details, e.g. clothing folds, and can have large variations of styles and topologies.

Various solutions have attempted to address the challenging problem of monocular garment reconstruction. Parametric methods [2, 23] represent a deformed garment in a human-driven manner, relying on the pose and shape of the underlying body. Clothing databases [37, 2] are often constructed by offline physical simulation or 3D scanning. These approaches [23, 10] can generate a garment model with similar global deformation to the image, but are difficult to obtain fine-grained details.

An alternative line of research [40, 41, 52] is to directly construct a pixel-aligned implicit field of clothed human from an input image. Such methods are difficult to obtain garments with reasonable structures and control the generation explicitly. For instance, methods based on signed distance fields (SDFs) [40, 41, 20] can only generate closed watertight surfaces. Implicit-based network [52] using unsigned distance fields are able to reconstruct open surfaces but could introduce additional artifacts when reconstructing mesh from the generated point clouds.

To tackle these limitations, we present *ImplicitPCA*, a hybrid framework that connects the good ends of both explicit mesh and implicit functions while being able to achieve physically plausible reconstructions. The core of our framework is a proposed parametric SDF network (PSDF Net) that bridges the PCA encoding of a parametric model and the implicit representation of high-fidelity reconstruction. To achieve this goal, we introduce two implicit modules: human SDF net and garment SDF net, which infer the occupancy field of the undressed human and 3D garment respectively from a single image. After these three modules are separately trained in a supervised manner, we leverage their representational advantages by closely coupling them in a novel optimization scheme at test time. In particular, to ensure the PSDF Net can generate a parametric mesh with high-quality geometric details that align with the input image, we impose a reconstruction loss to encourage the output of the PSDF network to be close to that of the garment SDF net. Additionally, to ensure physically plausible reconstruction, we introduce a collision-aware regression scheme with fast body-garment collision detection to avoid intersections between the outputs of human and garment SDF nets. Finally, to foster a faithful global deformation, we also propose a landmark constraint to encourage consistency between the 2D landmarks and the projections of pre-defined 3D landmarks on the generated mesh.

Unlike most parametric models that are generated from synthetic data, our PCA bases are extracted from MGN dataset [2] which are obtained from real scanned data. Hence, we are able to capture more realistic geometry details as shown in Figure 1. Furthermore, different from MGN, we are able to achieve more plausible reconstruction with geometry consistent with the input image and avoid human-garment collisions thanks to our hybrid framework. We evaluate our approach on a number of benchmarks. Experimental results demonstrate that our approach is superior to the state-of-the-art garment reconstruction methods [23, 10, 52] in terms of physic plausibility and the quality of geometry reconstruction.

In summary, our key contributions are the following three folds:

- For the realistic/flexible open garment surface reconstruction and generation, we introduce a hybrid parametric framework for 3D garment reconstruction, which can not only tightly collaborate explicit meshes and implicit fields but also jointly optimize the explicit garment meshes and implicit fields with the PCA parameters.
- We propose a novel collision-aware regress scheme for avoiding the collision, which is able to efficiently detect and optimize the cloth and body geometry in a differentiable fashion.
- We obtain state-of-the-art results on garment reconstruction from single images on the public benchmark MGN [2] and in-the-wild images in quantitative and qualitative evaluations.

2. Related Work

Parametric Models. Garment and human digitization from a single image is very challenging and requires geometry priors due to the ill-posed nature. And the garment and human contain more complex deformation to describe the detailed geometry, there are many researches focusing on the learning on deformation representation [44, 3, 50, 13, 18, 29]. Since human and garment geometry is shown to be

well reconstructed by PCA [31, 53], some parametric garment representations can generally be used to reconstruct the human and cloth and can be divided into human-based and garment-based methods.

Human-based models [42, 23, 30] represent the deformation of a garment mesh depending on the pose and shape of the underlying human body statistical models [1, 31, 28]. BCNet [23] introduces a layered garment on top of the SMPL model [31], as well as a set of learned skinning weights to improve the garment deformation. CAPE [30] uses a conditional Mesh-VAE-GAN with a mesh patch discriminator, to dress SMPL bodies with pose-dependent displacement cloth layers. Santesteban *et al*. [42] presents a data-driven framework that learns from the offline simulation to enable efficient virtual try-on with 250 fps. However, since the deformation of cloth heavily depends on the human shape and pose, these methods are able to generate the deformation of tight cloth, instead of loose cloth.

On the other hand, garment-based parametric methods [2, 37, 10] widely utilize the coarse-to-fine strategy to predict body shape and garment geometry jointly. For example, Bhatnagar *et al*. [2] present Multi-Garment Net (MGN), which predicts the global deformation and the detail displacement via two disentangled PCA coefficients of the pre-defined garment parametric models. Combining both human-based and garment-based methods, TailorNet [37] jointly learns a neural model, which achieves the detailed prediction of cloth by controlling the pose, shape, and style of template-based garments. Furthermore, to achieve a template-free method, SMPLicit [10] uses a parametric implicit function network with unsigned distance fields for garment generation with flexible topology. Parametric models provide well-defined geometry priors for garment reconstruction/generation. However, it is very challenging to reconstruct realistic and fine-grained garment details that are aligned with input images.

Neural Implicit Methods. The implicit representation has a great success in 3D shape representation and generation due to its flexible topology. For types of implicit functions, there are two main streams: binary signed distance function (SDF) [33, 38, 9, 15, 40, 11, 45] and continuous SDF [5, 34, 35, 24, 39] for closed surfaces, and unsigned distance function (UDF) [8, 47, 46, 52] for arbitrary shapes.

By using SDF for clothed human reconstruction, Saito *et al*. proposed PIFu [40] and PIFuHD [41], which are able to digitize highly detailed clothed humans with highly intricate shapes, hair styles, and texture in a unified way. The higher-fidelity performance is further raised by the higher resolution of input images and extra normal maps. Meanwhile, ARCH [21] learns a semantic deformation field using the parametric 3D body estimator to represent arbitrary shapes, but this method requires a body mesh as input, and is thus challenging to extend to garments with multiple components. Geo-PIFu [19] extends PIFu [40] by utilizing 3D information from a latent voxel representation to enrich the feature representation for a high-resolution reconstruction. However, Geo-PIFU is computationally intensive for both training and testing due to its adopted volume representation.

UDF is a powerful tool for cloth human reconstruction due to its ability to represent high-quality open surfaces. The pioneering works works NDF [8] and DUDE [47] use neural networks to predict unsigned distance fields to represent arbitrary surfaces without any closed surface data. They leverage the multi-scale techniques [8, 7] and normal vector field prediction [47] to enhance surface details. However, it is nontrivial and challenging to predict UDFs from an ambiguous 2D image for garment reconstruction, due to the lack of 3D space information. Different from the above approaches, AnchorUDF [52] optimizes the gradient direction of UDF via the extra set of 3D anchor points to obtain strong 3D context information for the distance fields.

Recently, there are some works focusing on novel 3D open surface representations, e.g., 3PSDF [4], WNF [6], HSDF [48]. Their tasks are reconstructing open surfaces with arbitrary topologies from sparse point clouds. As an application of 3PSDF [4], they show an example of reconstructed garments from a single-view image in T pose. While our method focus on various garment reconstruction tasks. The garments are often open surfaces but limited categories, and the topology is more restrict. Thus, a method relies on both explicit parametric models and implicit fields can get two good ends. Zhu *et al*. proposed an approach, called ReEF [54], to register a template mesh to semantic, shape and boundary implicit fields predicted from image by an optimization scheme. This method requires annotated datasets, well initialization and highly-correct correspondence for registration. Instead, we propose a hybrid representation that jointly collaborates the parametric models and flexible implicit fields via PCA parameters to predict layered 3D clothed human reconstruction with high-fidelity. Another important difference between our approach and the aforementioned methods is that the latter focuses only on cloth human reconstruction while our approach achieves a parameterized generation of clothed human via the PCA coefficients.

Body-Garment Collision detection & Response. Despite the great success of the above methods, all data-driven approaches have a critical weakness in handling body-garment collisions. Most methods [37, 17, 49, 16] commonly design additional loss terms to avoid and penalize geometric garment-body penetrations at training time, but such methods require extensive post-processing steps to fix collisions that occur during inference. To address this chal-

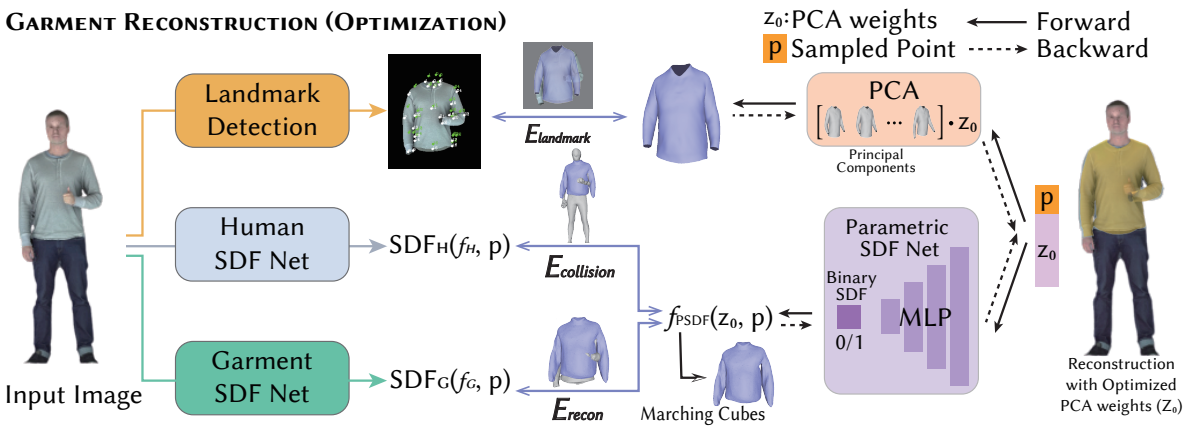


Figure 2: The architecture of our proposed garment reconstruction pipeline from a single-view image. It consists of three major modules: parametric SDF net, human SDF net, and garment SDF net. We train each network separately and use an optimization scheme to maintain high-fidelity detailed garments which avoid collisions with bodies.

lenge, Santesteban *et al*. [43] proposed a self-supervised loss, which is enabled by their powerful generative space of garment deformations. In contrast, we introduce a collision-aware regression network with fast body-garment collision detection and response, which is designed for the collision optimization between two SDFs of human and cloth. Our collision-aware network can be used to efficiently avoid collision in a differential fashion and also enhance the geometric details of garments according to human shape.

3. Methodology

Our framework consists of three major modules: parametric SDF net (PSDF Net), human SDF net, and garment SDF net. The PSDF Net consumes a 3D query point and a PCA latent code of the 3D clothing model and infers the binary occupancy of the reconstructed garment for the query point. The human and garment SDF nets take a single image as input and predict the occupancy of the undressed human and the garment respectively. At training time, each module is trained separately using the ground-truth data in a supervised manner. This is to ensure all the networks will learn the realistic shape priors. At test time, to obtain a high-fidelity and collision-free reconstruction, we introduce an optimization scheme that uses the output of the garment SDF net as the proxy of the main PSDF network to bridge the implicit representation and the parametric encoding. In particular, we use a reconstruction loss to encourage the outputs of the garment SDF net and the PSDF net to be close to each other. A novel collision loss is introduced to avoid the self-intersections between the predicted occupancy fields generated by the human and garment SDF nets. As the PSDF Net is fully differentiable, the self-supervised losses can provide gradient flows to optimize the latent PCA coefficients to obtain a high-quality and collision-free parametric encoding. In addition, a landmark loss is also proposed to foster the consistency between the 2D landmarks and the projections of the 3D landmarks pre-defined on the parametric mesh generated by the PCA encoding. Once the optimization converges, the parametric mesh represented by the optimized PCA code is used as our final reconstruction result.

3.1. Parametric SDF Net

We first introduce our Parametric SDF Net (named PSDF Net), which aims to learn the explicit mesh and implicit field jointly using a common space of PCA.

Given a set of N garment meshes $\mathcal{M} = \{M_1, M_2, \ldots, M_N\}$ with the same connectivity, we perform PCA [12] to obtain the first K principal components $\mathcal{B} = \{B_1, B_2, \ldots, B_K\}$ and the corresponding scores $\mathcal{S}$ for each garment mesh. The number of K is determined according to the percentage of variance (here we set it to 99%) to retain the geometry details while removing the noise. This formulation satisfies:

$$PCA(\mathcal{M}) \mapsto \{\mathcal{B}, \mathcal{S}\} \tag{1}$$

After that, each explicit garment with open surface can be represented as a linear combination of $\mathcal{B}$ with scores $S = (s_1, s_2, \ldots, s_K)$. Our PSDF Net aims to learn a mapping from the PCA scores $\mathcal{S}$ to the binary SDF (1 inside / 0 otherwise). Given the sample point $p = (x, y, z)$ near the surface and the PCA score $S = (s_1, s_2, \ldots, s_K)$, PSDF Net learns a function $f_{PSDF} : (p, S) \mapsto 0/1$, which aims to map (p, S) to an occupancy value 0/1, means outside/inside. Our network architecture is illustrated in Figure 2, which consists of multi-layer perceptrons to classify whether the sampled point p is an insider or outsider of the surface. Note that since garment meshes are often open surfaces, we enclose its open boundaries with the plane for sign distance

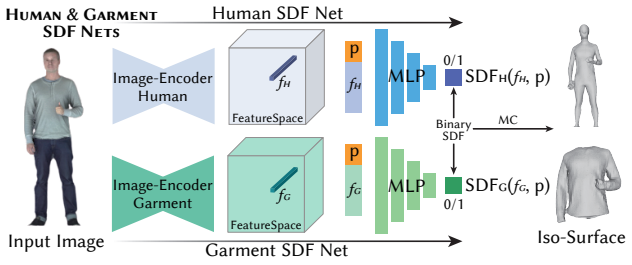


Figure 3: The architecture of PIFu-based Human and Garment Prediction Network. For each branch, we use the pixel-aligned features to predict the occupancy values and use MarchingCubes to extract the explicit meshes.

calculation [22] of sampled points. And the basis of PCA is extracted from the real scanned dataset – MGN [2], which contains lots of garments meshes with more rich and realistic geometric details.

We design a two-branch SDF prediction network for the garment and human reconstruction from a single image. Figure 3 presents the network architecture of the PIFu-based Human and Garment reconstruction. The detailed network architecture is adapted from the PIFu [40]. Given a single image I with a single person, a human parsing extractor [27] is applied to obtain the pixel-wise segmentation of garments labels automatically. Then, we pass them into different branches to predict their occupancy respectively, and finally the iso-surfaces are extracted by MarchingCubes [32]. Here, we model two functions, including Human SDF Prediction f_{human} and Garment SDF Prediction $f_{garment}$. These two functions take the pixel-wised feature and give 3D position X in the camera space as input, and predict the binary occupancy of the sampled point X. We formulate it as follows:

$$\begin{aligned} f_{human} &: (f_H, X) \mapsto 0/1 \\ f_{garment} &: (f_G, X) \mapsto 0/1 \end{aligned} \tag{2}$$

where the f_H and f_G are the pixel-wised feature of the segmented human image and segmented garment image via its corresponding image encoder, and 0/1 means that the sampled point is outside or inside. So, we can obtain the 3D human and 3D garments from an input single image I.

3.2. Collision-Aware Regression

The pixel-aligned approach is memory efficient and able to obtain high-fidelity 3D cloth and body geometries for the target subjects. However, directly predicting cloth and body implicit functions from an image may cause intersections between these two models since it is not collision-aware. As for some datasets, such as MGN [2], the ground truth data even is not body-garment collision-free.

We can obtain the SDFs $[SDF_G, SDF_H]$ of the garment and human by Eq. 2 in Sec. 3.1, we propose a collision-aware regression to optimize the SDF values of the garment and human to minimize the collision for the realistic and reasonable garment reconstruction.

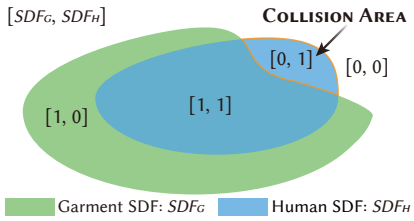


Figure 4: Body-garment collision detection via two signed distance fields. $[SDF_G(X) < 0.5, SDF_H(X) > 0.5]$ indicates a collision at sample X.

As shown in Figure 4, the impact zone of body-garment is easily detected by defining the collision condition of two implicit functions as:

$$\text{Collided at } X : SDF_G(X) < 0.5 \text{ and } SDF_H(X) > 0.5. \tag{3}$$

where the X is the position of the sampled points. If the point satisfies the above condition, the garment and human have collided at position X. Based on the observation, we design a novel collision penalty on the signed distance function to minimize the interpenetration between the SDF SDF_G of Garment and the underlying SDF SDF_H of Human. We formulate it as follows:

$$L_{collision} = \sum_{X \in \Omega} (SDF_G(X) - SDF_H(X)) T(SDF_G(X), SDF_H(X)) \tag{4}$$

$$T(x, y) = \begin{cases} 1, & \text{if } \quad x < 0.5 \quad \text{and} \quad y > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where the $T(\cdot)$ is a collision detection function. Ω is a set that contains sampled points in the space.

Different from previous methods which commonly search the nearest corresponding points on body shape for each point on the garment, our collision condition scheme effectively reduces the complexity of detecting collision points from $O(n^2)$ to $O(n)$. In Table 4, we show the efficiency comparison with BCNet [23] for the collision detection. From the table, we can clearly see that our method achieves fast speed and outperforms the SOTA – BCNet on collision detection by a large margin, which ensures that our framework can predict the 3D garments and humans from a single image effectively and efficiently.

3.3. Optimized Garment Reconstruction

Our goal is to reconstruct realistic and high-fidelity garments from a single-view image. However, it is very challenging to obtain ground-truth data since we need both real-world images and the corresponding garment meshes. Earlier works either build their own datasets with synthetic data

generated by physical simulation [23] or by using captured data with or without segmented labels [2, 51]. The former simulated dataset often cannot capture the details in the real environment. Meanwhile, the scanned dataset typically has limited garment variations as it is expensive and cumbersome to obtain large-scale 3D real data.

Hence, we introduce an optimized garment reconstruction pipeline based on the above modules with fixed training parameters without requiring a large number of training data. Formally, as illustrated in Figure 2, given an input image I and an initialized PCA weight $\mathbf{z_0}$ (*i.e.* the weight of PCA is randomly selected in training data), we can obtain the human SDF SDF_H, garment SDF SDF_G, initial garment mesh $M_{\mathbf{z_0}}$ and corresponding garment SDF $f_{PSDF}(\mathbf{z}, \cdot)$. To optimize the $\mathbf{z_0}$ to the optimal $\mathbf{z}$, we define the following target:

- $\boldsymbol{E}_{recon}$: mean square distance between parametric garment SDF $f_{PSDF}(\mathbf{z}, \cdot)$ and the garment SDF SDF_G, aims at reducing the difference between the predicted garments by the parametric SDF Net and the regressed garment implicit function by Garment SDF Net, *i.e.* $\boldsymbol{E}_{recon} = ||SDF_G(I, X) - f_{PSDF}(\mathbf{z}, X)||_2^2$;

- $\boldsymbol{E}_{collision}$: minimize the collision between the predicted garment SDF $f_{PSDF}(\mathbf{z}, \cdot)$ and human SDF SDF_H in Sec 3.2, *i.e.* $\boldsymbol{E}_{collision} = \sum_{X \in \Omega}(f_{PSDF}(\mathbf{z}, X) - SDF_H(I, X))T(SDF_H(I, X), f_{PSDF}(\mathbf{z}, X))$, where Ω is the set of sampled points;

- $\boldsymbol{E}_{reg}$: regularization term to penalize unrealistic reconstruction, *i.e.* $\boldsymbol{E}_{reg} = ||\mathbf{z}||_2^2$;

Apart from the above three energy terms, we also introduce the landmark loss to penalize landmark projection consistency error from the input image:

$$\boldsymbol{E}_{landmark} = \|\pi(LM_{3D}) - LM_{2D}\|_2^2 \tag{6}$$

where LM_{3D} represents the 3D landmarks on the mesh generated by PCA, the indices of the 3D landmarks are the same since we use a template garment. LM_{2D} represents 2D ground truth landmarks that are manually labeled. $\pi(\cdot)$ is a projection function with a given camera pose. The estimation of camera parameters is adopted from PIFu [40].

Finally, we can obtain the optimal PCA weight $\mathbf{z}$ by minimizing the formulation:

$$\boldsymbol{E}_{recon} + \omega_1 \boldsymbol{E}_{landmark} + \omega_2 \boldsymbol{E}_{collision} + \omega_3 \boldsymbol{E}_{reg} \tag{7}$$

where the $\omega_1, \omega_2, \omega_3$ are the weights to balance the optimization procedure. In our all experiments, we empirically set them as $\omega_1 = 1.0, \omega_2 = 0.1, \omega_3 = 5.0$.

4. Experiments & Evaluations

We first illustrate our network training and implementation details. Then we evaluate our proposed approach on three tasks: garment reconstruction from a single image, garment generation and interpolation via PCA parameters, where we show the superiority of our proposed methods on the large garment datasets from MGN [2], compare to the other strong baselines including SMPLicit [10], PIFu [40], BCNet [23], and AnchorUDF [52]. We further benchmark the performance of the garment reconstruction from the in-the-wild images. In the end, ablation studies are conducted in order to validate our key designs.

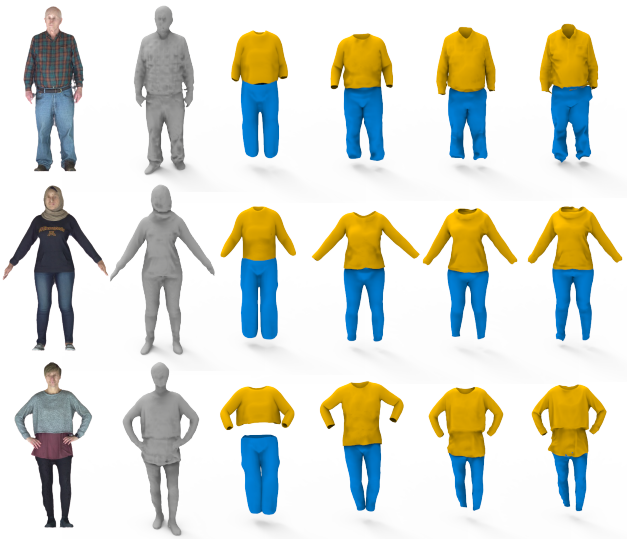
4.1. Implementation Details

Datasets. We primarily use MGN [2] for the majority of our experiments. The MGN dataset contains five categories garments and 154 textured garment models. For training, we adapt the same method as PIFu and rendered 360-degree images of each garment along the yaw axis and obtained 48240 images in total. For the landmark on the 2D image, we use the same setting as the DeepFashion2 [14]. The data-splitting are followed from AnchorUDF [52].

Implementation. For our network, we first train on the PSDF Net and Human/Garment SDF Nets simultaneously. Our PSDF Net is used to map the parametric explicit garment mesh and the corresponding SDF into one common space of PCA parameters. Human/Garment SDF Nets aims to predict the human SDF and garment SDF from a single image. After that, we fix all the trainable parameters, and optimize the PCA parameters $\mathbf{z}_0$ to by minimizing all the introduced energies in Sec. 3.3 of our main paper. Similar to DeepFashion3D [53], we focus on front-view reconstruction with single person. Our whole network is implemented in PyTorch [36]. All the occupancy prediction network is implemented with MLPs, and the image-encoder for human/garment is adapt from PIFu [40]. For the training, the whole network is optimized by Adam optimizer [26], we set batch size as 8 and use a learning rate that is from 0.0001 and decays by 0.9 every 1000 steps. The PSDF Net will converge at 2000 epochs, and Human/Garment SDF Nets takes 200 epochs to converge. Empirically, our network converges after 1 day. Table 1 shows the detailed runtime statistics of Garment/Human SDF Net for once image inference, and the iterative optimization costing of the Parametric SDF Net. The optimization stage for garment reconstruction a single image takes 400 epochs to converge. Note that all the experiments are evaluated on a computer with an i7-7700K CPU, 64G RAM, and a GeForce GTX 2080Ti graphic card.

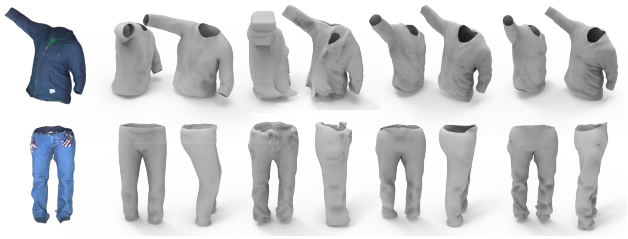
Table 1: The timing (second) of each component in the testing stages, averaging in different cloth categories on the MGN dataset [2]. Notice that the timing of Parametric SDF Net is the total 400 epoch optimization time-cost.

Components	Garment SDF Net	Human SDF Net	Parametric SDF Net
Timing	0.01	0.01	4.52



(a) Input (b) PIFu (c) SMPLicit (d) BCNet (e) Ours (f) GT

Figure 5: Cloth reconstruction comparison on single-view images with two implicit function methods PIFu [40]) and SMPLicit [10], and a mesh template-based method BCNet [23].



(a) Input (b) BCNet (c) AnchorUDF (d) Ours (e) GT

Figure 6: Shape reconstruction comparison with the baseline methods (BCNet [23], AnchorUDF [52]).

4.2. Garment Reconstruction

We qualitatively and quantitatively compare our method to other state-of-the-art for single-view garment reconstructions. Two metrics [52] including the Chamfer Distance (CD) and average point-to-surface Euclidean distance (P2S) are used for qualitative evaluation. Table 2 presents the reconstruction errors for each method. Benefiting from implicitly-proxied parametric encoding, our method can optimize the poses and shapes of humans and garment details, thus achieving the best performance for garment and human reconstruction.

Table 2: Chamfer Distance (CD) and Point to Surface (P2S) errors ($\times 10^{-3}$) of garment and human by different single-view reconstruction methods on MGN dataset. '-' means that AnchorUDF only reconstructs the garment mesh. Lower is better.

Methods	Garment		Human	
	CD↓	P2S↓	CD↓	P2S↓
BCNet [23]	4.053	4.512	3.808	4.310
SMPLicit [10]	9.012	10.591	3.724	4.179
AnchorUDF [52]	0.635	0.762	-	-
Ours	**0.494**	**0.172**	**0.332**	**0.364**

Figure 5 presents the qualitative comparisons on the garment reconstruction from a single-view image. Our model can successfully capture the detailed wrinkles from the input image and achieve realistic garment reconstruction. It is clear to see that SMPLicit [10] and BCNet [23] fail to capture the detailed wrinkles shown in the image. Although PIFu [40] are able to output pixel-aligned high-resolution results, they are not able to separate human and cloth. In Figure 6, AnchorUDF [52] predicts the cloth as open surface and enhances the details. Since they are not topology-aware, the side-view image shows clear artifacts.

(a) Input (b) PIFu (c) SMPLicit (d) BCNet (e) Ours

Figure 7: Garment reconstruction comparison on in-the-wild images (PIFu [40], SMPLicit [10]), BCNet [23].

We further evaluate the performance on in-the-wild images compared with three state-of-the-art methods, BCNet [23], SMPLicit [10], and PIFu [40]. We use the tool [25] in SMPLicit to remove background and a human parsing extractor [27] to obtain a pixel-wise segmentation of clothing labels (*i.e.* upper clothes, pants), automatically.

As shown in Figure 7, the alternative baselines yields got realistic results, but they could hardly reconstruct the detailed garments, such as the wrinkles on the garments. We observe that our approach outperforms the alternative approaches and achieves better performance in terms of visual quality. In particular, the garments predicted by our method are more realistic. For various poses, we also conduct an experiment trained on the publicly available TailorNet [37] dataset which maintains simulated garments on quite different human poses. We tested on more in-the-wild images with various poses. As shown in Fig. 8, our method outperforms compared works in both human-pose alignment and garment detail prediction.

(a) Input (b) PIFu (c) SMPLicit (d) BCNet (e) Ours

Figure 8: Cloth reconstruction comparison on in-the-wild images under arbitrary poses with PIFu [40]), SMPLicit [10] and BCNet [23].

Discussion ImplicitPCA v.s. ReEF Both our method and ReEF [54] are approaches combining explicit and implicit 3D representation of garments for single-view reconstruction. ReEF [54] register a template mesh to semantic, shape and boundary implicit fields predicted from image by an optimization scheme. There are two difficulties to get high-quality ReEF results. Firstly, learning each implicit field from images requires annotating large customized dataset as they mentioned in the paper. For instance, professional artists annotate the garment boundaries on the scan surfaces and may link the incomplete boundary segments into smoothed closed curves with their expertise in garments' shape. Secondly, the quality of mesh registration is depending on the initialization and the correspondence between source and target. The initialization uses 2D pose estimation as guidance for SMPL prediction, which may fail to align with the shape implicit field in hard cases, e.g., different human poses. And the shape fitting is only processed in the subset of mesh vertices, which is a sparse correspondence. Instead of registering, our method jointly utilize the good ends of explicit and implicit representation to achieve high-fidelity and physically-plausible results. Explicit mesh representation has clear boundary information and topology which is used for 2D landmark supervision. Implicit function captures pixel-aligned features from images efficiently which is used for garment detail reconstruction. This hybrid parametric model enables realistic and topology-correct reconstructions, see more results in Fig. 8.

4.3. Garment Generation & Interpolation

Figure 9 presents the garment generation via the explicit PCA parameters. In this figure, we show the visual results by controlling the first two dimensions of the PCA parameters. It is clearly observed that the first two dimensional PCA parameters can control the orientation and size of the garments. So, equipped with the explicit PCA parameters, our framework enables the novel generation of garments with a given factor. Table 3 shows the number of PCA components for different cloth in our experiments.

Table 3: The number of PCA components for different cloth categories on the MGN dataset [2].

Cloth Categories	T-shirt	Shirt	Short-Pants	Pants
# PCA components	35	20	13	20

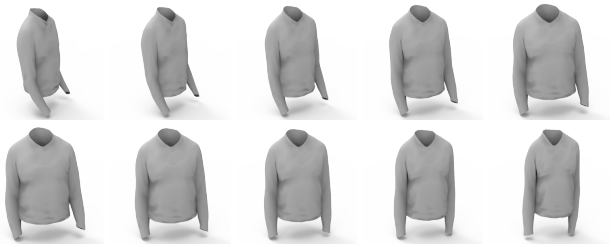


Figure 9: Top: effect of the first principal component (controlling orientation). Bottom: effect of the second principal component (controlling size).

Since our framework bridges the explicit parametric model and implicit fields via PCA parameters, we can get parametric implicit proxies. In particular, garment generation and interpolation naturally are achieved with the help of explicit PCA parameters. In Figure 10 we show some generated garments by randomly sampling on the PCA parameters of training data. The results demonstrate that our

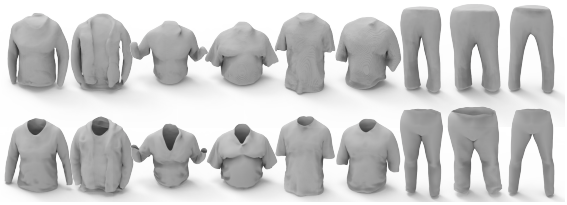


Figure 10: Random generation of the shirt, T-shirt, and pants shapes, where the training data is from MGN dataset [2]. The first row is the extracted iso-surfaces via MC, and the second row is the corresponding meshes.

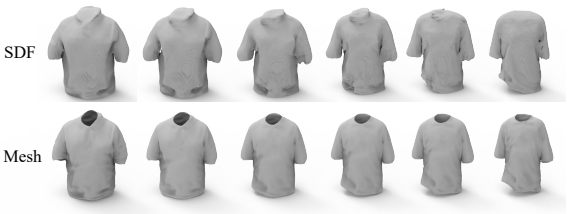


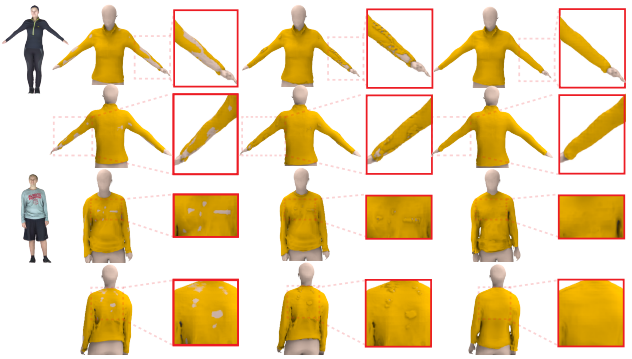
Figure 11: Shape interpolation of T-shirts represented by SDF and mesh using PCA weights in our method. The first and last columns are two input shapes.

model is able to generate diversified 3D garments in both parametric meshes and learned implicit fields.

The interpolation results are further evaluated, which are shown in Figure 11. We interpolate the PCA parameters from two input T-shirts linearly to control the generated implicit fields of the garment explicitly. The first row shows the generated results by PSDF that are represented as closed meshes extracted via MarchingCubes [32]. The second row is the corresponding open surfaces for each interpolated parameter. Our approach achieves smooth and continuous interpolation with detailed wrinkles. These experiments demonstrate the capabilities of our PSDF Net for achieving high-fidelity inference.

4.4. Ablation Studies

With *vs*. without collision loss. To evaluate the advantages of our collision-aware module proposed in 3.2, we separately show the impact of with and without collision loss on the garment and human reconstruction. In Figure 12(c), it is observed that our method successfully minimizes the collision between the garment and body shape in Figure 12(a). Besides, our method can prevent collisions on the unseen back of the garments, as well as enhance the geometric details of the garments concerning the body shape. There are also some post-processing algorithms (*e.g.* [23]) to handle the collision, but it can introduce some bulges artifacts (see Figure 12(b)) by moving the vertex along the normal direction. Furthermore, our collision detection method is 10^4 times faster than the post-processing approach, which benefited from the parametric implicit proxies (see Table 4).

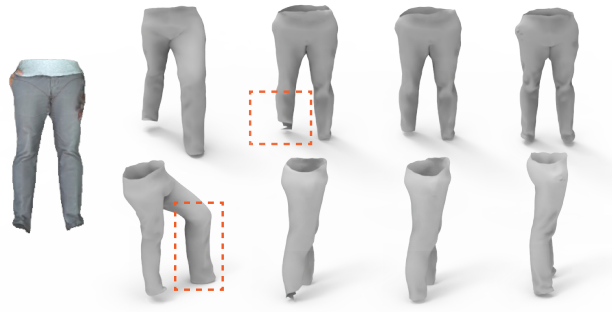


(a) w/o $\boldsymbol{L}_{collision}$ (b) post-process (c) w/ $\boldsymbol{L}_{collision}$

Figure 12: Ablation of collision loss for regression of cloth and human SDF from a single-view image.

Table 4: Timing of different collision detection methods. G_p and H_p ($\times 10^3$) are the number of garment and human points and t(s) is the time cost.

Method	G_p / H_p / t	G_p / H_p / t	G_p / H_p / t
BCNet	6/6/0.492	19/22/2.609	181/384/44.38
Ours	7/7/**5.10**$\times 10^{-5}$	20/20/**7.70**$\times 10^{-5}$	400/400/**8.39**$\times 10^{-5}$



(a) (b) (c) (d) (e)

Figure 13: Ablation of each energy term in the optimization routine: (a) input image; (b) $\boldsymbol{E}_{landmark}$; (c) $\boldsymbol{E}_{landmark} + \boldsymbol{E}_{recon}$; (d) $\boldsymbol{E}_{all}$ means the summation of all the optimized energies: $\boldsymbol{E}_{recon} + \boldsymbol{E}_{landmark} + \boldsymbol{E}_{collision} + \boldsymbol{E}_{reg}$; (e) Ground Truth.

Table 5: Chamfer Distance (CD) and Point to Surface (P2S) errors ($\times 10^{-3}$) of garment optimization schemes with different energy terms on MGN dataset. Lower is better.

Methods	$\boldsymbol{E}_{landmark}$	$\boldsymbol{E}_{landmark} + \boldsymbol{E}_{recon}$	$\boldsymbol{E}_{all}$
CD $\downarrow$	23.6	0.624	0.494
P2S $\downarrow$	12.9	0.242	0.172

With *vs*. without each term in optimization routine. Figure 13 illustrates the results generated by different set-

tings of terms in the optimization routine. Each setting contains the regularization loss term to avoid instability synthesis. Although the model is constrained with only landmark loss term $\boldsymbol{E}_{landmark}$ generates 3D shapes aligned well with 2D key points, it leads to posing ambiguity without considering depth information. Furthermore, the model trained with both landmark and SDF reconstruction loss terms $\boldsymbol{E}_{landmark}+\boldsymbol{E}_{recon}$ generates a trouser in better gestures with more wrinkle details. However, the bottom of the trouser has apparent bending artifacts. This is due to the ignoring of collision. The model with full constraints outperforms others in generating detailed garments with desired gestures closer to the ground truth. The landmark-, reconstruction- and collision constraints jointly improve the quality of high-resolution cloth reconstruction. Moreover, by comparing Chamfer distance (CD) and point-to-surface (P2S) in Table 5, we observe that the best performances are achieved only when all energies are fused.

(a) Input Image (b) Reconstruction

Figure 14: A failure case with cloth collision and occlusion on the 2D image.

5. Conclusions & Discussions

In this paper, we have presented a novel hybrid representation for 3D cloth and humans, called IMPLICITPCA. We integrate the parametric-based method and implicit-based method and introduce an collision-aware regression network with fast collision detection and penalization. Our experiment indicates that high-fidelity garments following physical rules can be inferred from a monocular image. We envision that IMPLICITPCA would benefit a series of Metaverse applications, such as virtual try-on and garment editing.

Finally, although our model achieves state-of-the-art performance in realistic and high-fidelity garment reconstruction, our method still has some failure cases. As shown in Figure 14, some poses with cloth collisions result in a lack of information, thus making it difficult to infer the side-view of the person. These cases can be very challenging for our collision-aware module, and the reconstructed cloth will crash into the person as shown in Figure 14. Hence, for future work, some priors of poses can be applied to enhance the quality of garments during extremely challenging cases. The generalization of our method in garment categories is also limited by the training dataset. For those garments categories that are not covered in the training set, the recovery process can only converge to the most similar category. We will both extend dataset and explore a typology augmentation method in the future.

Acknowledgments

This work was supported by CCF-Tencent Open Fund, the Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (No. JQ21013), the National Natural Science Foundation of China (No. 62061136007 and No. 61872440), the Royal Society Newton Advanced Fellowship (No. NAF\R2\192151) and Open Research Projects of Zhejiang Lab (No. 2021KE0AB06).

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape completion and animation of people. *ACM Transactions on Graphics (TOG)*, 24(3):408–416, 2005. 3

[2] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, oct 2019. 2, 3, 5, 6, 7, 8, 9

[3] S.-Y. Chen, L. Gao, Y.-K. Lai, and S. Xia. Rigidity controllable as-rigid-as-possible shape deformation. *Graphical Models*, 91:13–21, 2017. 2

[4] W. Chen, C. Lin, W. Li, and B. Yang. 3psdf: Three-pole signed distance function for learning surfaces with arbitrary topologies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18522–18531, 2022. 3

[5] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 3

[6] C. Chi and S. Song. Garmentnets: Category-level pose estimation for garments via canonical space shape completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3324–3333, 2021. 3

[7] J. Chibane, T. Alldieck, and G. Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2020. 3

[8] J. Chibane, A. Mir, and G. Pons-Moll. Neural unsigned distance fields for implicit function learning. *arXiv preprint arXiv:2010.13938*, 2020. 3

[9] J. Chibane and G. Pons-Moll. Implicit feature networks for texture completion from partial 3d data. In *European Conference on Computer Vision*, pages 717–725. Springer, 2020. 3

[10] E. Corona, A. Pumarola, G. Alenya, G. Pons-Moll, and F. Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11875–11885, 2021. 2, 3, 6, 7, 8

[11] B. Deng, J. P. Lewis, T. Jeruzalski, G. Pons-Moll, G. Hinton, M. Norouzi, and A. Tagliasacchi. Nasa neural articulated shape approximation. In *European conference on computer vision*, pages 612–628. Springer, 2020. 3

[12] K. P. F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. 4

[13] L. Gao, G. Zhang, and Y. Lai. L p shape deformation. *Science China Information Sciences*, 55(5):983–993, 2012. 2

[14] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5337–5345, 2019. 6

[15] K. Genova, F. Cole, A. Sud, A. Sarna, and T. Funkhouser. Deep structured implicit functions. *arXiv preprint arXiv:1912.06126*, 2, 2019. 3

[16] E. Gundogdu, V. Constantin, S. Parashar, A. S. Banadkooki, M. Dang, M. Salzmann, and P. Fua. Garnet++: Improving fast and accurate static 3d cloth draping by curvature loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[17] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8739–8748, 2019. 3

[18] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[19] T. He, J. Collomosse, H. Jin, and S. Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *arXiv preprint arXiv:2006.08072*, 2020. 3

[20] T. He, Y. Xu, S. Saito, S. Soatto, and T. Tung. Arch++: Animation-ready clothed human reconstruction revisited. *ArXiv*, abs/2108.07845, 2021. 2

[21] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung. Arch: Animatable reconstruction of clothed humans. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3090–3099, 2020. 3

[22] A. Jacobson, D. Panozzo, et al. libigl: A simple C++ geometry processing library, 2018. https://libigl.github.io/. 5

[23] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, and H. Bao. Bcnet: Learning body and cloth shape from a single image, 2020. 1, 2, 3, 5, 6, 7, 8, 9

[24] C. Jiang, A. Sud, A. Makadia, J. Huang, M. Nießner, T. Funkhouser, et al. Local implicit grid representations for 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020. 3

[25] Kaleido AI GmbH. Remove background, 2018. https://www.remove.bg/. 7

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] P. Li, Y. Xu, Y. Wei, and Y. Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5, 7

[28] S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2848–2856, 2015. 3

[29] X.-J. Li, J. Yang, and F.-L. Zhang. Laplacian mesh transformer: Dual attention and topology aware network for 3d mesh classification and segmentation. In *European Conference on Computer Vision*, pages 541–560. Springer, 2022. 2

[30] Z. Li, M. Oskarsson, and A. Heyden. Learning to implicitly represent 3d human body from multi-scale features and multi-view images. In *International Conference on Pattern Recognition (ICPR)*, pages 8968–8975. IEEE, 2021. 3

[31] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 3

[32] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 5, 9

[33] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 3

[34] M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktashmotlagh, and A. Eriksson. Deep level sets: Implicit surface representations for 3d shape inference. *arXiv preprint arXiv:1901.06802*, 2019. 3

[35] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 3

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[37] C. Patel, Z. Liao, and G. Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7365–7375, 2020. 2, 3, 8

[38] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger. Convolutional occupancy networks. In *European conference on computer vision*, pages 523–540. Springer, 2020. 3

[39] E. Remelli, A. Lukoianov, S. Richter, B. Guillard, T. Bagautdinov, P. Baque, and P. Fua. Meshsdf: Differentiable iso-surface extraction. *Advances in Neural Information Processing Systems*, 33:22468–22478, 2020. 3

[40] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 1, 2, 3, 5, 6, 7, 8

[41] S. Saito, T. Simon, J. M. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d

human digitization. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 81–90, 2020. 2, 3

[42] I. Santesteban, M. A. Otaduy, and D. Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, volume 38, pages 355–366. Wiley Online Library, 2019. 3

[43] I. Santesteban, N. Thuerey, M. A. Otaduy, and D. Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11763–11773, 2021. 4

[44] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *EUROGRAPHICS/SIGGRAPH Symposium on Geometry Processing*, pages 109–116, 2007. 2

[45] J.-H. Tang, W. Chen, J. Yang, B. Wang, S. Liu, B. Yang, and L. Gao. Octfield: Hierarchical implicit functions for 3d modeling. *arXiv preprint arXiv:2111.01067*, 2021. 3

[46] R. Venkatesh, T. Karmali, S. Sharma, A. Ghosh, R. V. Babu, L. A. Jeni, and M. Singh. Deep implicit surface point prediction networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12653–12662, 2021. 3

[47] R. Venkatesh, S. Sharma, A. Ghosh, L. Jeni, and M. Singh. Dude: Deep unsigned distance embeddings for hi-fidelity representation of complex 3d surfaces. *arXiv preprint arXiv:2011.02570*, 2020. 3

[48] L. Wang, J. Yang, W. Chen, X. Meng, B. Yang, J. Li, and L. Gao. Hsdf: Hybrid sign and distance field for modeling surfaces with arbitrary topologies. In *Advances in Neural Information Processing Systems*. 3

[49] T. Y. Wang, D. Ceylan, J. Popovic, and N. J. Mitra. Learning a shared shape space for multimodal garment design. *arXiv preprint arXiv:1806.11335*, 2018. 3

[50] Y.-J. Yuan, Y.-K. Lai, J. Yang, Q. Duan, H. Fu, and L. Gao. Mesh variational autoencoders with edge contraction pooling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 274–275, 2020. 2

[51] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 6

[52] F. Zhao, W. Wang, S. Liao, and L. Shao. Learning anchored unsigned distance functions with gradient direction alignment for single-view garment reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12674–12683, 2021. 2, 3, 6, 7

[53] H. Zhu, Y. Cao, H. Jin, W. Chen, D. Du, Z. Wang, S. Cui, and X. Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *European Conference on Computer Vision*, pages 512–530. Springer, 2020. 3, 6

[54] H. Zhu, L. Qiu, Y. Qiu, and X. Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3845–3854, 2022. 3, 8