

Layered-Garment Net: Generating Multiple Implicit Garment Layers from a Single Image^{*}

Alakh Aggarwal¹, Jikai Wang¹, Steven Hogue¹, Saifeng Ni², Madhukar Budagavi², and Xiaohu Guo¹

¹ The University of Texas at Dallas, Richardson, TX, US
{[alakh.aggarwal](mailto:alakh.aggarwal@utdallas.edu), [jikai.wang](mailto:jikai.wang@utdallas.edu), [ditzley](mailto:ditzley@utdallas.edu), [xguo](mailto:xguo@utdallas.edu)}@utdallas.edu

² Samsung Research America, Plano, TX, US
{[saifeng.ni](mailto:saifeng.ni@samsung.com), [m.budagavi](mailto:m.budagavi@samsung.com)}@samsung.com

Abstract. Recent research works have focused on generating human models and garments from their 2D images. However, state-of-the-art researches focus either on only a single layer of the garment on a human model or on generating multiple garment layers without any guarantee of the intersection-free geometric relationship between them. In reality, people wear multiple layers of garments in their daily life, where an inner layer of garment could be partially covered by an outer one. In this paper, we try to address this multi-layer modeling problem and propose the Layered-Garment Net (LGN) that is capable of generating intersection-free multiple layers of garments defined by implicit function fields over the body surface, given the person’s near front-view image. With a special design of garment indication fields (GIF), we can enforce an implicit covering relationship between the signed distance fields (SDF) of different layers to avoid self-intersections among different garment surfaces and the human body. Experiments demonstrate the strength of our proposed LGN framework in generating multi-layer garments as compared to state-of-the-art methods. To the best of our knowledge, LGN is the first research work to generate intersection-free multiple layers of garments on the human body from a single image.

Keywords: Image-based Reconstruction · Multi-layered Garments · Neural Implicit Functions · Intersection-free.

1 Introduction

Extracting 3D garments from visual data such as images enables the generation of digital wardrobe datasets for the clothing and fashion industry, and is useful in Virtual Try-On applications. With the limitation on certain classes of garments, it is already possible to generate explicit upper and lower garment meshes from a single image or multi-view images [1, 2], to introduce different styles to the garments, such as length, along with varying poses and shapes [3–5], and to transfer the garments from one subject to another [1].

^{*} Aggarwal, Wang, Hogue, and Guo are partially supported by National Science Foundation (OAC-2007661).

However, to the best of our knowledge, none of the existing approaches have the capability of generating multiple *intersection-free* layers of clothing on a base human model where an inner layer of garment could be partially covered by an outer one without any intersection or protrusion. This does not conform to reality because people wear multiple layers of garments in their daily life. The existing techniques either generate a single layer of upper-body cloth (e.g., T-shirt, jacket, etc.) and a single layer of lower-body cloth (e.g., pants, shorts, etc.) without any overlap in their covering regions [2, 1], or generate multiple garment layers, but without any guarantee on their intersection-free geometry [5].

The fundamental challenge here is to ensure intersection-free between multiple garment layers when they overlap. Existing approaches to garment representation are based on *explicit* models, by using either displacement fields over SMPL surface (SMPL+D) [6, 1] or skinned meshes on top of SMPL [2]. However, with explicit mesh representations, it is very difficult to ensure intersection-free between multiple garment layers. SMPLicit [5] is an implicit approach that generated multiple layers of garments but does not handle intersections among multiple layers. In this paper, we propose to use a set of *implicit* functions – signed distance fields (SDF), to represent different layers of garments. The benefit is that the intersection-free condition can be easily enforced by requiring the SDF of the inner layer to be greater than the SDF of the outer one. We call this the *Implicit Covering Relationship* (Sec. 3.1) for modeling multi-layer garments.

There are two challenges associated with the such implicit representation of garments as well as the enforcement of implicit covering relationship: (1) Most of the garments are *open* surfaces with boundaries, while SDF can represent *closed* surfaces only. (2) The implicit covering relationship should only be enforced in those regions where two layers overlap, but how can we define such overlapping regions? In this paper, we solve these two challenges by proposing an implicit function called *Garment Indication Field* (GIF, Sec. 3.2) which successfully identifies those regions where the garment has “holes” – the open regions where the garment does not cover. With such garment indication fields, we not only can enforce the implicit covering relationships between layers but also can extract the open meshes of garments by trimming the closed marching cubes surfaces.

We propose a *Layered-Garment Net* (LGN), which consists of a parallel SDF subsystem and GIF subsystem, that can take an image of the person as input, and output the corresponding SDF and GIF for each garment layer. Specifically, based on the projection of the query point in image space, we obtain its local image features from the encoded features given by a fully convolutional encoder. Using the local image features and other spatial features of the query point, we train different decoder networks for different layers of garments to predict their SDF and GIF, respectively. The network is trained end-to-end, utilizing a covering inconsistency loss given by GIFs and SDFs of different layers, along with other loss functions to regress the predictions to the ground-truth values. The contributions of this paper can be summarized as follows:

- We present a Layered-Garment Net, the first method that can model and generate multiple intersection-free layers of garments and the human body, from a single image.
- We enforce an implicit covering relationship among different layers of garments by using multiple signed distance fields to represent different layers, which guarantees that multiple layers of garments are intersection-free on their overlapping regions.
- We design garment indication fields that can be used to identify the open regions where the garments do not cover, which can be used to identify the overlapping regions between different layers of garments, as well as to extract open meshes of garments out of the closed surfaces defined by SDF.

2 Related Works

In this section, we will review the recent works in two areas of research that are related to our work. We consider **Full Human Body Reconstruction** where the focus is on generating a good quality clothed human model and **Individual Garment Surface Reconstruction** where the focus is on obtaining individual garments for a human model.

Full Human Body Reconstruction Many recent works generated explicit representations of human body mesh using parametric models for naked human models to handle varying geometry [7–9]. This allows them to modify the shape and pose of the generated model according to shape parameters β and pose parameters θ . The underlying idea is to obtain the parameters β and θ that closely defines the target human body, and apply linear blend skinning using the blend shapes and blending weights to generate the final human body geometry. Bogo et al. [10] obtained these parameters and fitted a human body model from single unconstrained images. Many deep learning-based methods [11, 12] have since then come up, that estimate the shape and pose parameters of a human model. Smith et al. [13] employed the use of silhouettes from different viewpoints to generate the human body. Subsequently, some research works [14, 15] also used semantic segmentation of human parts to ensure more accurate parameter estimation. However, the above-mentioned works only generate naked human models and do not reconstruct clothed human models.

To address this issue, several recent research works have focused on the displacements of a naked body. Alldieck et al. [6] used frames at some continuous interval from a video of a subject rotating in front of the camera to ensure accurate parameter estimation from different viewpoints and used SMPL+D for clothed human body reconstruction. Such SMPL+D representation uses a displacement vector for the vertices of the naked human body model to represent clothing details and was later used for single image reconstruction [16]. Tex2Shape [17] was able to obtain better displacement details by predicting the displacement map for a model that aligns with the texture map of the model. Several recent works [18] generate explicit dynamic human models. However, since all the above methods are only based on a naked human body model, they cannot generate a human body wearing complex garments like skirts, dresses,

long hair, etc. To address these issues, some research works [19, 20] used a volumetric representation of the human body with voxelized grids. Ma et al. [21] obtain the point clouds of clothed humans with varying garment topology. Some recent works [22, 23] also focus on generative approaches for 3D clothed human reconstruction.

There have been some recent works focusing on the implicit clothed body surface representation. Mescheder et al. [24] used the occupancy field to determine if a point is inside or outside a surface of any object from the ShapeNet [25] dataset, and then used a classifier to generate a surface dividing the 3D space into inside/outside occupancy values. They calculated occupancy values for each point of the voxel grid and used marching cubes [26] to generate the surface. They do not have to store voxel grid representation or any other mesh information for all the data instances. Different from the occupancy field, Chibane et al. [27] predicted an unsigned distance field using a neural network, and projected the points back to the surface to generate a point cloud-based surface using the gradient of the distance field at that point, and could be used to further generate a complete mesh surface. Several recent works [28–30] predicted a Signed Distance Field and used marching cubes [26] to generate a mesh surface. This ensures more accurate geometry because of the implicit field’s dependence on distance. Based on the above works on implicit fields, PIFu [31], PIFuHD [32], StereoPIFu [33], GeoPIFu [34], PaMIR [35] take a 2D image or depth data of human as input, and after extracting the local encoded image features for a point, they predict the occupancy field of the dressed body. MetaAvatar [36] represent cloth-specific neural SDFs for clothed human body reconstruction. Other recent works [37–39] aim to dynamically handle the reconstruction of animatable clothed human models via implicit representation. Several other works [40, 5, 41, 42] also use implicit fields for 3D human reconstruction. Bhatnagar et al. [43] combine use base explicitly defined SMPL model to implicitly register scans and point clouds. The method identifies the region between garment and body, however, it does not reconstruct different garment layers. Handling individual garment regions like Garment Indication Field is more complicated. Scanimate [42] reconstructs a dynamic human model and utilizes an implicit field for fine-tuning their reconstruction. Instead of supervision, they utilize Implicit Geometric Regularization [44] to reconstruct surfaces using implicit SDF in a semi-supervised approach.

Individual Garment Surface Reconstruction Instead of simply generating a human body model with displacements, Multi-Garment Net (MGN) [1] generates an explicit representation of parametric garment models with SMPL+D. Using single or multiple images, it predicts different upper and lower garments that are parameterized for varying shapes and poses. However, MGN cannot produce garments that do not comply with naked human models, like skirts and dresses. TailorNet [3] uses the wardrobe dataset from MGN and applies different style transforms like sleeve-length to obtain different styles of garments. DeepCloth [4] enable deep-learning based styling of garments. SIZER [45] provides a dataset enabling resizing of the garment on the human body. Deep Fashion3D [46] generates a wardrobe dataset, consisting of complex garment shapes like skirts and

dressess. BCNet [2] uses a deformable mesh on top of SMPL to represent garments and proposes a skinning weights generating network for the garments to support garments with different topologies. SMPLicit [5] obtains shape and style features for each garment layer from the image and uses these parameters to obtain multiple layers of garments, and uses a distance threshold to reconstruct overlapping garment layers. However, they do not guarantee intersection-free reconstruction. GarmentNets [47] reconstructs dynamic garments utilizing Generalized Winding Number [48] for occupancy and correct trimming of openings in garment meshes. Their approach, however, does not provide a garment’s indicator field.

To the best of our knowledge, none of the existing works can generate overlapping intersection-free multiple layers of garments where an inner layer could be partially covered by an outer one. All existing works on individual garment generation [1, 3, 46, 2] use explicit mesh representation, making them difficult to ensure intersection-free between different layers. SMPLicit [5] does not guarantee intersection-free reconstruction among different layers, especially in overlapping regions. In this paper, we resort to implicit representation and model the multiple layers of garments with signed distance fields (SDF) which makes it easy to enforce the implicit covering relationship among different layers of garment surfaces with the help of a carefully designed implicit garment indication field (GIF). The combination of these two implicit functions, SDF and GIF, makes the modeling and learning of multi-layer intersection-free garments possible.

3 The Method

Given a near-front-facing image of a posed human, we aim to generate the different intersection-free garment surface layers. The reconstructed surfaces should follow a covering relationship between each other and the body. Our proposed Layered-Garment Net (LGN) can generate implicit functions of Signed Distance Field (SDF) and Garment Indication Field (GIF) for different layers of garments over varying shapes and poses. An overview of our approach is given in Fig. 1.

3.1 Implicit Covering Relationship

For two layers of garments i and j , let layer i be partially covered by layer j . If a point p belongs to their overlapping regions, the SDF values $s_i(p)$ and $s_j(p)$ for the two layers should follow the covering relationship:

$$s_j(p) < s_i(p). \quad (1)$$

This is illustrated in Fig. 2(left). The inequality does not hold for all the points in 3D space but only holds for the overlapping region between the two layers. We are only interested in the points near the surface of the garment layer. Let us consider an example where layer i is a pant and layer j is a shirt. The inequality Eq. (1) should not be satisfied in the leg region of the human body, otherwise, this would result in the generation of a shirt layer on top of the pant layer in the leg region, where the shirt originally does not exist. This problem is shown in Fig. 2(right). Hence, we need an indicator function for both layers, and only ensure that the implicit covering relationship Eq. (1) holds on points that are related to both layers i and j . We call this indicator function for layer i the *Garment Indication Field* (GIF), and denote it as $h_i(p)$ (Sec 3.2).

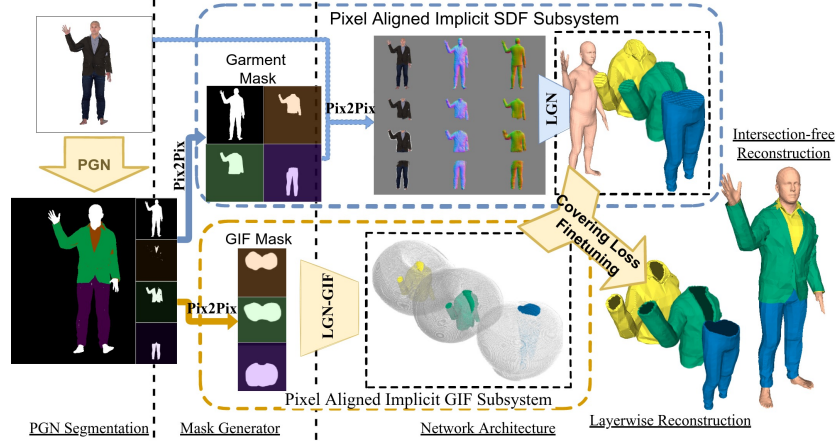


Fig. 1. Given an input image, our method first obtains PGN [49] segmentation and incomplete masks for each garment from segmentation. Then the complete masks of garments are generated by Pix2Pix-HD Garment Mask generator. Similarly the indicator masks for garments are generated by Pix2Pix-HD Indicator Mask generator. Using the masked input image, the front and back normals are obtained using Pix2Pix-HD Normal Subnetwork. Then an Encoder and Decoders of LGN network use masked images and normals and predicts SDF value $s_i(p)$ of layer i for any point p in 3D space. LGN-GIF further uses Indicator Masks ind_i to obtain GIF value $h_i(p)$ of layer i for any point p in 3D space. Finally, LGN is fine-tuned with the covering loss in Eq. (2) to avoid intersection among different layers.

To ensure the network’s SDF predictions follow the Implicit Covering Relationship inequality for relevant points p , we can define the covering loss for all layers of surfaces for our network as follows:

$$\mathcal{L}_{cov}(p) = \sum_{j=1}^N \sum_{i \in C(j)} h_j(p) * h_i(p) * [\max(s_j(p) - s_i(p), 0) + \lambda(s_j(p) - s_i(p))^2], \quad (2)$$

where $C(j)$ is the set of layers partially covered by layer j . The multiplication with $h_j(p)$ and $h_i(p)$ guarantees that the covering loss only applies to the points in the overlapping region between two layers. The last term regularizes the difference between the two SDFs. We choose $\lambda = 0.2$ in all our experiments.

3.2 Garment Indication Field

For a garment and a query point p , we use its generalized winding number [48], denoted as $W(p)$, to distinguish the open regions from the regions concerned with garment surfaces. Since all the garments are open surfaces, $W(p)$ is equal to 0.5 at the opening regions. $W(p) > 0.5$ for a point inside the surface, and keeps increasing as the point gets farther inside. Similarly, $W(p) < 0.5$ for a point outside the surface, and keeps decreasing as it goes further away from the open regions. In far-off regions and outside the surfaces, $W(p) \leq 0$. Using

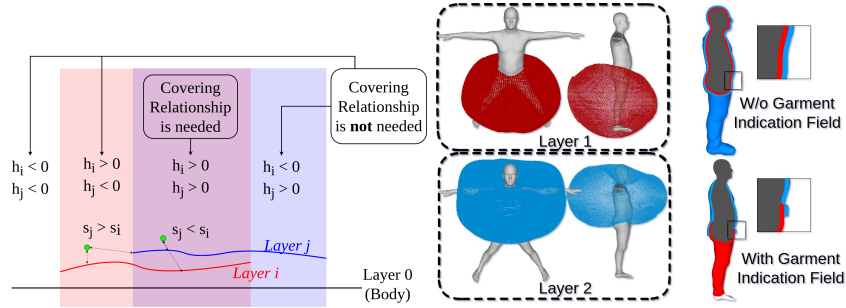


Fig. 2. (Left) For a point p associated with two layers of surfaces i and j , where layer i is partially covered by layer j , it should satisfy $s_i(p) > s_j(p)$ in their overlapping region where $h_i(p) > 0$ and $h_j(p) > 0$. For other regions, this relationship may not satisfy. (Right) Garment Indication Fields (GIF) of an inner layer pant and an outer layer shirt are used to constrain the covering relationship only in their overlapping region. Without GIF, the outer layer would completely cover the inner layer since $s_i(p) > s_j(p)$ would be enforced everywhere.

different field functions as a function of winding number, we can have different observations as shown in Fig. 3.

Observation 1: $o(p) = W(p) - 0.5$ gives the occupancy field for a garment. This has been shown in Fig. 3(b). We call $o(p)$ the *winding occupancy*. This helps us in obtaining the sign of SDF for a non-watertight mesh. Since all garments, in particular, are non-watertight open mesh, for any query point p in 3D space, the distance $d(p)$ to its nearest surface point is essentially an *unsigned* distance because there is no inside/outside for the open surface. Thus we use $o(p)$ to obtain a watertight surface mesh with marching cubes first, then compute a *ground-truth SDF* $s'(p)$ for the watertight garment surface.

Observation 2: $h'(p) = W(p) * (W(p) - w_h)$ gives an indication field of the garment opening region, where $0.5 < w_h < 1$. As previously discussed, $W(p)$ is greater than 0.5 inside the opening region and the surface mesh, and it keeps on increasing inside the mesh. Similarly, $W(p)$ is less than 0.5 outside the opening region and keeps decreasing away from the region outside the mesh. In this paper, we choose $w_h = 0.75$ for all garments. For any point that is inside the mesh and away from the garment opening region, $W(p) > 0.75$, so $h'(p)$ is positive. Similarly, if it is outside the mesh and away from the garment opening region, $W(p) < 0$, so $h'(p)$ is positive too. However, for any point that is located close to the 0.5-level isosurface, $0 < W(p) < 0.75$, so $h'(p)$ is negative. In this way, $h'(p)$ indicates the open region of the garment. This can be observed from Fig. 3(c).

Furthermore, it also follows that $h'(p) - \delta$, for some $\delta \rightarrow 0^+$ gives a bound region of the garment closer to the mesh. This has been shown in Fig. 3(d). We observe that, for $\delta = 0.01$, we get a good quality bound for this indication field. Thus we define the following function as Garment Indication Field (GIF) for the

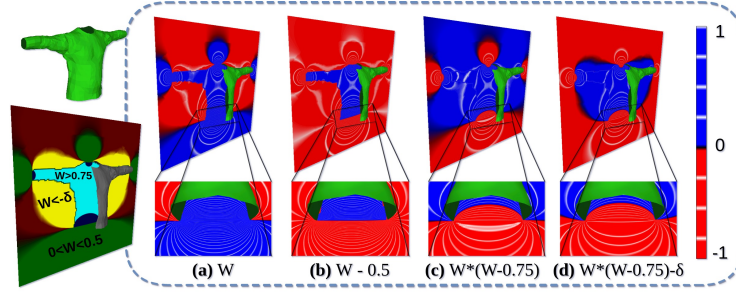


Fig. 3. For a given garment mesh (non-watertight), we show the cross-section views of the following fields: (a) Generalized Winding Number W , (b) Winding Occupancy $W - 0.5$, (d) “Hole” Region Indication $W * (W - 0.75)$, and (e) Garment Indication Field $W * (W - 0.75) - \delta$. Transition from (c) to (d) allows concise bound for GIF, which will not intersect with nearby body surfaces.

garment surface:

$$\hat{h}(p) = (\text{sign}[W(p) * (W(p) - 0.75) - \delta] + 1) * 0.5. \quad (3)$$

Here $\hat{h}(p) = 1$ means the point is in the region close to the garment surface, otherwise $\hat{h}(p) = 0$. Such ground-truth GIF values will be used for enforcing the covering relationship in Eq. (2).

3.3 Layered-Garment Net

Given an input image I of a person, we first obtain the garment segmentation P on the image using Part-Grouping Network [49]. It is possible to obtain different garment masks g_i for garment layer i using the corresponding pixel color. However, the mask g_i may not be complete because of overlap with outer layers. Hence, we train a Garment Mask generator, that takes the incomplete mask g_i and PGN segmentation P as input, and outputs a corrected garment mask g'_i .

Like PiFU-HD [32], we follow a similar pipeline, however, with no requirement for a Fine-Level network, but only a Coarse network for each garment layer. We also use semi-supervised Implicit Geometric Regularization (IGR) [44] for fine-tuning SDF prediction on the surface. For layer i , we use mask g'_i on input image I and using a Normal Subnetwork masked with g'_i , we obtain front and back normals [32]. Let’s call the concatenation of masked input image and masked front and back normals for layer i as N_i .

LGN consists of a common SDF Encoder for all layers, that gives feature encoding for N_i as F_i . For a given point p in 3D space, we obtain a local pixel feature by orthogonal projection $\pi(p)$ of p on F_i , and barycentric interpolation. For the point p , we also obtain spatial features like depth. Using the spatial features and local pixel-aligned features, the layer’s SDF Decoder $\mathcal{S}_i(\cdot)$ predicts SDF $s_i(p)$.

Similarly, to identify if point p lies in the garment region, we also obtain Indicator Mask ind_i of layer i from PGN P and incomplete mask g_i by training

an Indicator Mask generator. Then we train a common GIF Encoder that gives encoding F'_i and the layer's GIF Decoder $\mathcal{H}_i(\cdot)$ to obtain GIF value for layer i as $h_i(p)$.

Mask generators for each garment follow the same architecture as front and back Normal Subnetworks in [32], i.e. Pix2PixHD network [50]. We have a common SDF Encoder and front and back Normal Subnetworks among all garment layers and body layers. A common GIF Encoder is defined for all garment layers. However, we separately define SDF Decoders, GIF Decoders, Garment Mask generators, and Indicator Mask generators for each garment - shirt, pant, coat, skirt, dress.

$$s_i(p) = \mathcal{S}_i(F_i(\pi(p)), \phi(p)), \quad h_i(p) = \mathcal{H}_i(F'_i(\pi(p)), \phi(p)), \quad (4)$$

where the spatial feature $\phi(p)$ here is depth.

The L1 loss for the generated SDF is formulated as the following L^1 norm:

$$\mathcal{L}_{sdf}(p) = \sum_{i=0}^N |s_i(p) - \hat{s}_i(p)|, \quad (5)$$

where $\hat{s}_i(p)$ is the ground-truth SDF value for point p from layer i , and N is the number of garment layers.

Similarly, the L1 loss for the predicted GIF is formulated as follows:

$$\mathcal{L}_{gif}(p) = \sum_{i=1}^N |h_i(p) - \hat{h}_i(p)|, \quad (6)$$

where $\hat{h}_i(p)$ is the ground-truth GIF for the garment layer i and query point p .

We fine-tune network parameters for SDF prediction using Implicit Geometric Regularization (IGR) [44]. Loss for IGR is given as follows:

$$\begin{aligned} \mathcal{L}_{igr}(p) &= \tau(p)\ell_{\mathcal{X}}(p) + \lambda(\|\nabla_p s_i(p)\| - 1)^2, \\ \ell_{\mathcal{X}}(p) &= |s_i(p)| + \|\nabla_p s_i(p) - n_p\|, \end{aligned} \quad (7)$$

where $s_i(p)$ is the SDF value at p , $\tau(p)$ is an indicator of a point on surface \mathcal{X} and n_p is the surface normal at point p .

3.4 Training and Inference

We first pre-train the Garment Mask generator and Indicator Mask generators on PGN segmentation and incomplete mask as inputs for each garment category - shirt, pant, coat, skirt, dress. To train the network, we sample 20,480 points on the surface of each layer. We add normal perturbation $\mathcal{N}(0, \sigma = 5cm)$ on these points to generate the near-surface samples. We then add random points in 3D space using a ratio of 1 : 16 for the randomly sampled points w.r.t. the near-surface samples. These sampled points are used to optimize the SDF prediction of all garment layers and covering loss between each layer of the garment. We similarly sample points from the 0.5 level iso-surface of the ground-truth Garment Indication Fields (GIFs) of each layer garment layer and add normal perturbation $\mathcal{N}(0, \sigma = 5cm)$ on these points to generate garment indications. We add random points in 3D space using a ratio of 1 : 16 for the randomly sam-

pled points w.r.t. the garment indicating samples. For GIFs, we add additional points along the edges of ground truth mesh to obtain accurate trimming.

Given an input image I , we first obtain PGN segmentation image P which contains different garments in the image. For each garment, we obtain their incomplete masks g_i . Using P and g_i for each layer, we obtain garment masks g'_i and ind_i . We leave out the indicator mask prediction for the body layer since it is not required to obtain GIF for the body. It is assumed that the GIF value for the body layer is 1 at any point. For all the near-surface sampled points, we calculate the ground-truth SDF values \hat{s}_i for each layer i as explained in 3.2. The encoder and decoder are warmed-up by training with the loss \mathcal{L}_{sdf} and \mathcal{L}_{igr} as defined in Eq. (5) and Eq. (7). We also calculate ground-truth GIF values \hat{h}_i for each layer $i > 0$ with the loss \mathcal{L}_{gif} as defined in Eq. (3). Using the predicted h_i values and the predicted s_i values for all the sampled points, the network is fine-tuned with the covering loss as defined in Eq. (2). This ensures that the output SDF values follow the covering relationship inequality as defined in Eq. (1). For all the garments indicating sampled points, we calculate their ground-truth GIF values for each layer of the garment. Using the ground-truth GIF values for each layer, the GIF value prediction is optimized.

During the inference, after obtaining the SDF values for each layer, we use marching cubes to obtain its triangle mesh. Then, we apply a trivial post-processing step using predictions, to update SDF values. For a given point p where GIF of both layers i and j overlap: if $s_j > s_i - \epsilon$, $s_j = s_i - \epsilon$ where ϵ is a very small number. Experimentally, we use ϵ to be $1e-3$. Finally, all the triangular meshes obtained for each layer are trimmed by the predicted GIF values on the vertices of the mesh. To trim the garment opening regions, the triangles which have different signs of GIF values for its three vertices are selected, and the triangle is trimmed by linearly interpolating GIF values over each edge. Thus, we finally obtain multiple layers of garments along with the reconstructed Layer-0 body that follows the covering relationship.

For both training and inference, we rely on covering the relationship manually specified with the input image. Different garment layers are then obtained from the output of LGN by satisfying the covering relationship.

4 Experiments

4.1 Dataset and Implementation Details

Dataset Preparation. Our multi-layer garment dataset is constructed from 140 purchased rigged human models from AXYZ [51]. For each rigged model, we first perform SMPL [7] fitting to obtain its body shape and pose parameters. We generate eight images from different views for each human model and run semantic segmentation on each image with Part Grouping Network (PGN) [49]. Using the fitted SMPL, we obtain those segmentations on the SMPL surface and map them to the UV texture space of SMPL. This enables us to perform texture stitching [52] to generate the segmentation texture map. By projecting the texture segmentation onto the 3D human model, we obtain the segmentation of different 3D garment meshes, followed by minor manual corrections on some garment boundaries. Our processed garments include the categories of Shirt,

Coat, Dress, Pant (long and short), and Skirt, while Shirt/Coat/Dress all contain three subcategories of no-sleeve, short-sleeve, and long-sleeve. Detailed statistics of the processed garments are provided in the supplementary document.

	Model	<i>P2S</i>		Model	<i>P2S</i>		Model	<i>Chamfer</i>	<i>P2S</i>
A(i)	BCNet	9.75	A(ii)	BCNet	3.84	B	PiFU-HD	1.22	1.19
	SMPLicit	9.12		SMPLicit	6.01		BCNet	1.93	1.96
	Ours	9.09		Ours	4.04		Ours	2.75	2.6

Table 1. Comparison results (in *cm*) for **A** per-garment Point-to-Surface on (i) Digital Wardrobe [1], (ii) SIZER [45], and **B** Full body reconstruction on BUFF Dataset [53]

Using different garments, we synthesize around 12,000 different combinations of multi-layer garments on top of a layer-0 SMPL body, in 7 different poses. When combining different garment types, we follow the assumption that the length of sleeves for the inner layer should NOT be shorter than that of the outer layer. Otherwise, the sleeves of the inner layer are covered by the outer garment and there is no visual clue to tell its length. We then use this combination of generated garment models with a layer-0 body to train our LGN. We use the synthesized combinations of multi-layer garments as the training set. The geometries of garments are corrected to make sure no intersections exist and the different layers of garment follow the covering relationship. For testing, we use BUFF [53] and Digital Wardrobe [1] datasets. The dataset preparation details are discussed in the supplementary document.


Implementation Details. The base architecture of our LGN is similar to that of PiFu [31] and PiFu-HD [32] since they also predict implicit fields aligned with image features. For SDF Subsystem, we first obtain garment masks from PGN segmentation of input image using Garment Mask Generators and obtain masked front and back normals using Normal Subnetworks, which are Pix2Pix-HD [50] networks. We use 4 stacks of Stacked Hourglass Network (HGN)[54] to encode the image features from the concatenation of normals and image. From spatial features from points and local encoded features by performing bi-linear interpolation of projected points on image feature space, different Multi-layer Perceptron (MLP) decoder layers predict SDF values for each layer of the garment, with layer 0 being the human body. Similarly, for GIF Subsystem, Indicator Mask Generators obtain GIF masks. 4 stacked-HGN encodes image features from the concatenation of PGN segmentation and indicator mask. Then, GIF is predicted using GIF Decoders. To optimize the network, we first pre-train Mask Generators. Then we individually train SDF Subsystem and GIF Subsystems. Thereafter, we use the covering loss to fine-tune SDF prediction to avoid the intersection, and GIF to ensure appropriate trimming of the open region on garment surfaces, and a consistent multi-layer covering relationship. We evaluate our methods and test with various state-of-the-art approaches on mainly two areas - 3D Clothed Human Reconstruction and Individual Garment Reconstruction. The quantitative comparison of 3D Clothed Human Reconstruction of our method with BCNet [2], PiFU-HD [32] and SMPLicit [5] are shown in the

supplementary document. We omit comparison with MGN [1], Octopus [52] and PiFU [31] because of the availability of better reconstruction methods.

4.2 Quantitative Comparisons

We compare our methods with the state-of-the-art (Tab. 1) approaches on three publicly available datasets: Digital Wardrobe Dataset [1], SIZER Dataset [45] and BUFF Dataset [53]. We use Digital Wardrobe Dataset and SIZER Dataset to compare individual garment reconstructions, and BUFF Dataset [53] to compare full human body reconstruction. It is to be noted that since the datasets mentioned consist of only 2 layers of garments – upper and lower, we cannot make a comparison with them on multi-layer garment reconstruction. Please also note, we do not use BCNet [2] data set to have a fair comparison with BCNet. Also, we are unable to compare our results with DeepFashion3D [46] data set because the dataset only consists of garments and no human body.

Method	Im1	Im2	Im3	Im4	Im5
SMPLicit	12.9	32.7	1.67	21.3	18.3
Ours	0.34	0	0	0.16	0



Im1
Im2
Im3
Im4
Im5

Fig. 4. Penetration depths (in cm).

Individual Garment Reconstruction To compare our method with the state-of-the-art garment reconstruction approaches [5, 2], we select 96 models from Digital Wardrobe Dataset [1] and 97 models from SIZER Dataset [45]. We use segmented Upper and Lower garments available with Dataset for comparison. We calculate the Mean P2S Error per garment between reconstructed garments and their ground-truth counterpart and observe the performance of our approach with other approaches in Tab. 1 A(i)&(ii). Our model outperforms state-of-the-arts on Digital Wardrobe Dataset [1]. For SIZER [45], BCNet performs better due to assuming reconstruction of 2 layer garments only, since segmentation for 3 layer of garments does not exist in the data set.

In Fig. 4, we calculate the Maximum Penetration Depth between different reconstructed garment layers and make a comparison with SMPLicit [5]. It can be seen that our work outperforms the state-of-the-art in this case.

Full Human Body Reconstruction We show in Tab. 1 B the comparison of our method with the state-of-the-arts on full human body reconstruction, on 26 models consisting of different subjects and clothes from BUFF Dataset [53]. We calculate Chamfer distance and Point-to-surface (P2S) error between ground-truth human models and reconstructed full body surface. We do not compare with SMPLicit, because they have no method for full body reconstruction. Please



Fig. 5. Cover Loss Finetuning

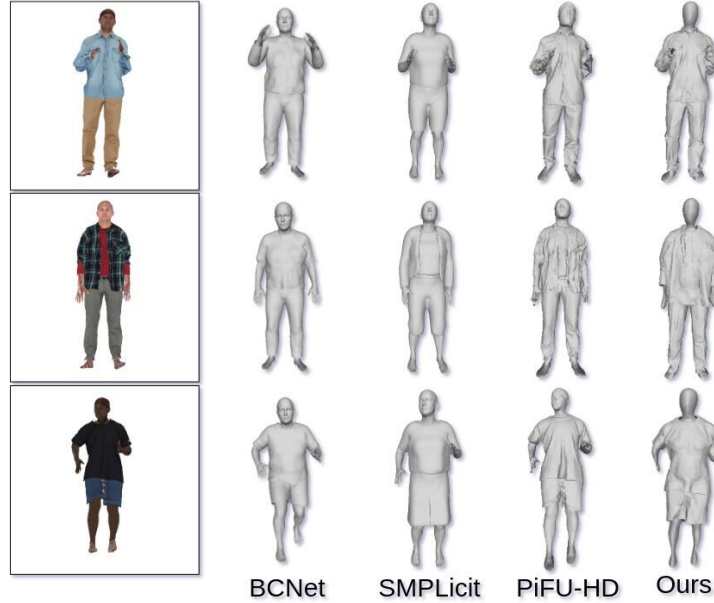


Fig. 6. From left to right, qualitative comparison of full-body reconstruction on 3D clothed human from ground truth (left), and results from BCNet [2], SMPLicit [5], PiFu-HD [32] and Ours (LGN).

note that we encounter lower results in this case than state-of-the-arts because we do not focus on accurate naked body (layer 0) reconstruction.

4.3 Qualitative Results

We compare the reconstruction quality of garment surfaces on the human body in Fig. 6. We can observe that our method (LGN) reconstructs a more detailed 3D human body than state-of-the-art explicit model reconstruction methods like BCNet, showing the effectiveness of implicit model reconstruction in comparison explicit approach. Also, SMPLicit generates a very coarse structure and loses many finer details for the clothes on the human body. Since we can generate individual implicit garment surfaces, we can retain finer details, especially between different layers. Since our networks are fine-tuned with IGR Loss, we reconstruct garments of similar quality to PiFu-HD without using fine-level networks.

In Fig. 7, we further show different challenges faced in the reconstruction of different garment surfaces. In the top row, we show the effect of covering relationship on multiple layers of garment reconstruction, specifically for the Shirt and Pant layers. From the given image, we expect the Pant layer to cover the Shirt layer without intersections. However, BCNet generates Shirt covering Pants, according to their pre-defined template. On the other hand, SMPLicit completely misses covering relation. In the bottom row, we show the reconstruction of the Coat layer above the Shirt layer as in the image. We expect two layers of gar-

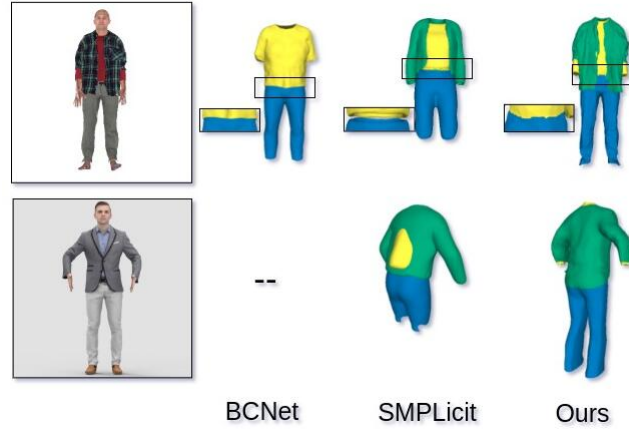


Fig. 7. Comparison between the reconstruction results of ours, BCNet [2] and SMPLicit [5]. Our reconstruction achieves intersection-free between different layers by satisfying the implicit covering relationship, while BCNet cannot reconstruct multi-layered garment structure, and the result from SMPLicit does not have such guarantee and has clear intersections between different layers.

ment reconstruction for the upper body. However, since BCNet is based on an explicit reconstruction of garments based on a displacement map on the SMPL body, it cannot reconstruct two-layer geometry for the upper body. Since SMPLicit does not guarantee intersection-free reconstruction, we find intersections between Shirt and Coat layer. Since the results of our LGN satisfy the covering relationship in Eq. (1), we get the expected output of garments in both cases.

In Fig. 5, we show how Covering Loss affects the reconstruction output. Without covering loss finetuning, inner layers intersect with outer layers.

5 Conclusion, Limitation, and Future Work

We introduce a novel deep learning-based approach that reconstructs multiple non-intersecting layers of garment surfaces from an image. Our approach enforces the implicit covering relationship between different garment layers and the human body and identifies overlapping regions of different garment layers, as well as extract open (non-watertight) meshes. To the best of our knowledge, Layered-Garment Net (LGN) is the first approach that can handle the intersection-free reconstruction of multiple layers of garments from a single image.

Our approach currently does not handle color information, since obtaining good texture for multiple reconstructed layers is difficult. Other neural implicit functions (e.g. Neural Radiance Fields) can address this issue. Our approach does not handle more challenging geometries consisting of manifold garment surfaces and details like pockets, hoodies, collars, etc., and some challenging poses, like limbs close to the body, etc. Also, since the naked human body model was not the focus of this work, the current approach does not handle the detailed full-body reconstruction. These issues can be a major improvement for future work.

References

1. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: Proceedings of the IEEE/CVF international conference on computer vision. (2019) 5420–5430
2. Jiang, B., Zhang, J., Hong, Y., Luo, J., Liu, L., Bao, H.: Bcnet: Learning body and cloth shape from a single image. In: European Conference on Computer Vision, Springer (2020) 18–35
3. Patel, C., Liao, Z., Pons-Moll, G.: 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. (2020) 7365–7375
4. Su, Z., Yu, T., Wang, Y., Li, Y., Liu, Y.: Deepcloth: Neural garment representation for shape and style editing. arXiv preprint arXiv:2011.14619 (2020)
5. Corona, E., Pumarola, A., Alenya, G., Pons-Moll, G., Moreno-Noguer, F.: Smplicit: Topology-aware generative model for clothed people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 11875–11885
6. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8387–8397
7. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34** (2015) 1–16
8. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: 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. (2019) 10975–10985
9. Osman, A.A., Bolkart, T., Black, M.J.: Star: Sparse trained articulated human body regressor. In: European Conference on Computer Vision, Springer (2020) 598–613
10. Bogu, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European conference on computer vision, Springer (2016) 561–578
11. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7122–7131
12. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 459–468
13. Smith, B.M., Chari, V., Agrawal, A., Rehg, J.M., Sever, R.: Towards accurate 3d human body reconstruction from silhouettes. In: 2019 International Conference on 3D Vision (3DV), IEEE (2019) 279–288
14. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV), IEEE (2018) 484–494
15. Lassner, C., Romero, J., Kiefel, M., Bogu, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 6050–6059
16. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single rgb camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 1175–1186

17. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 2293–2303
18. Su, Z., Wan, W., Yu, T., Liu, L., Fang, L., Wang, W., Liu, Y.: Mulaycap: Multi-layer human performance capture using a monocular video camera. *arXiv preprint arXiv:2004.05815* (2020)
19. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3d human body shapes. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018) 20–36
20. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 7739–7749
21. Ma, Q., Saito, S., Yang, J., Tang, S., Black, M.J.: 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*. (2021) 16082–16093
22. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020) 6469–6478
23. Wang, L., Zhao, X., Yu, T., Wang, S., Liu, Y.: Normalgan: Learning detailed 3d human from a single rgb-d image. In: *ECCV*. (2020)
24. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2019) 4460–4470
25. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015)
26. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics* **21** (1987) 163–169
27. Chibane, J., Mir, A., Pons-Moll, G.: Neural unsigned distance fields for implicit function learning. *arXiv preprint arXiv:2010.13938* (2020)
28. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2019) 165–174
29. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *arXiv preprint arXiv:1905.10711* (2019)
30. Sitzmann, V., Chan, E.R., Tucker, R., Snavely, N., Wetzstein, G.: MetaSDF: Meta-learning signed distance functions. *arXiv preprint arXiv:2006.09662* (2020)
31. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 2304–2314
32. Saito, S., Simon, T., Saragih, J., Joo, H.: 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*. (2020) 84–93

33. Hong, Y., Zhang, J., Jiang, B., Guo, Y., Liu, L., Bao, H.: Stereopifu: Depth aware clothed human digitization via stereo vision. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2021) 535–545
34. He, T., Collomosse, J., Jin, H., Soatto, S.: Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *arXiv preprint arXiv:2006.08072* (2020)
35. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021)
36. Wang, S., Mihajlovic, M., Ma, Q., Geiger, A., Tang, S.: Metaavatar: Learning animatable clothed human models from few depth images. In: *Advances in Neural Information Processing Systems (NeurIPS)*. (2021)
37. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: Arch: Animatable reconstruction of clothed humans. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020) 3093–3102
38. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2021) 9054–9063
39. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2021) 14314–14323
40. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. *arXiv preprint arXiv:2106.02019* (2021)
41. Shao, R., Zhang, H., Zhang, H., Cao, Y., Yu, T., Liu, Y.: Doublefield: Bridging the neural surface and radiance fields for high-fidelity human rendering. *arXiv preprint arXiv:2106.03798* (2021)
42. Saito, S., Yang, J., Ma, Q., Black, M.J.: Scanimate: Weakly supervised learning of skinned clothed avatar networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2021) 2886–2897
43. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer (2020) 311–329
44. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099* (2020)
45. Tiwari, G., Bhatnagar, B.L., Tung, T., Pons-Moll, G.: Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In: *European Conference on Computer Vision (ECCV)*, Springer (2020)
46. Zhu, H., Cao, Y., Jin, H., Chen, W., Du, D., Wang, Z., Cui, S., Han, X.: Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In: *European Conference on Computer Vision*, Springer (2020) 512–530
47. Chi, C., Song, S.: Garmentnets: Category-level pose estimation for garments via canonical space shape completion. *arXiv preprint arXiv:2104.05177* (2021)
48. Jacobson, A., Kavan, L., Sorkine-Hornung, O.: Robust inside-outside segmentation using generalized winding numbers. *ACM Transactions on Graphics (TOG)* **32** (2013) 1–12
49. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018) 770–785

50. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8798–8807
51. : (Axyz design 3d people models and character animation software <https://secure.axyz-design.com/><https://secure.axyz-design.com/>)
52. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: 2018 International Conference on 3D Vision (3DV), IEEE (2018) 98–109
53. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
54. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision, Springer (2016) 483–499