DOI: 10.1111/cgf.14792

COMPUTER GRAPHICS *forum*
Volume 42 (2023), number 6, e14792

Feature Representation for High-resolution Clothed Human Reconstruction

Juncheng Pu,[1] Li Liu,[1,2] Xiaodong Fu,[1,2] Zhuo Su,[3] Lijun Liu[1,2] and Wei Peng[1,2]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China
1409110293@qq.com,xiaodong_fu@hotmail.com,cloneiq@126.com,weipeng1980@gmail.com
[2]Computer Technology Application Key Laboratory of Yunnan Province, Kunming, China
[3]The School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
suzhuo3@mail.sysu.edu.cn

Abstract

Detailed and accurate feature representation is essential for high-resolution reconstruction of clothed human. Herein we introduce a unified feature representation for clothed human reconstruction, which can adapt to changeable posture and various clothing details. The whole method can be divided into two parts: the human shape feature representation and the details feature representation. Specifically, we firstly combine the voxel feature learned from semantic voxel with the pixel feature from input image as an implicit representation for human shape. Then, the details feature mixed with the clothed layer feature and the normal feature is used to guide the multi-layer perceptron to capture geometric surface details. The key difference from existing methods is that we use the clothing semantics to infer clothed layer information, and further restore the layer details with geometric height. We qualitative and quantitative experience results demonstrate that proposed method outperforms existing methods in terms of handling limb swing and clothing details. Our method provides a new solution for clothed human reconstruction with high-resolution details (style, wrinkles and clothed layers), and has good potential in three-dimensional virtual try-on and digital characters.

Keywords: modelling; geometric modelling, modelling; implicit surfaces, modelling; surface reconstruction

CCS Concepts: • Computing methodologies → Computer graphics; Shape modelling; Mesh models

1. Introduction

The high-fidelity 3D model of clothed human is crucial in many graphic related applications, including virtual reality, digital human, and virtual try-on, and so on, which requires the reconstruction of clothed human with arbitrary postures and clothing details (style, wrinkles and clothed layers) from details and accurate feature representation. Considering the monocular images are more available than camera matrix [ZWG*21], many images-based methods using high-capacity deep learning models have been proposed for the clothed human reconstruction recently. For instance, Zhu et al. [ZZW*19] proposed a template-based method that recovers detailed human shapes from even a single image with joint, anchor and per-vertex. Although the parametric template-based methods can perform the human posture, the template is mainly used to infer the naked body, it is often difficult to deal with the topological changes caused by clothing details. Instead, the implicit representation have shown great promise in acquiring reconstructions of clothed human with multifarious details. As representative of the state-of-the-art approaches, the pixel-aligned implicit function (PIFu) [SHN*19] has been proposed to infer clothed human shape, which performs implicit function prediction from latent code and the z value of query point from a single image. In addition, a subsequent study by Saito et al. [SSSJ20] enhances the geometric results of PIFu for fine-grained details recovery by using predicted normal priors and higher resolution input. However, PIFu-related methods are difficult to fit the challenging posture, and they are prone to generate faulty structures, for example broken or slim limbs. Consequently, above methods concentrate on the reconstruction of 3D posture or the recovery of clothing details and they are limited to the dimensional constraints of pixel information, failing to generate 3D model with both arbitrary posture and clothing details.

Researcher has also proposed depth-aware reconstruction methods to overcome the limitations about challenging posture mentioned above. He et al. [HCJS20] predicted the occupancy probability from pixel-aligned and voxel-aligned feature. However, the application scenario and robustness of this method are limited by the depth features extracted from known 3D meshes. Zheng et al. [ZYLD21] estimated parameter model from a single image and convert it into 3D feature volume. Then, the voxel-aligned feature is extracted from the 3D feature volume and predicts the occupancy probability for query point. Using stereo images, Yang et al. [HZJ*21] can infer human shape and predict human location in camera space without normalizing human shape into a canonical space. Previous methods have been limited by the large variation in human pose that cannot be adequately represented by parametric or volumetric shapes. Different from these methods, we render the depth semantic code onto the voxelization of skinned multi-person linear model (SMPL) [LMR*15]. Compared with mesh or simple voxel, depth priors provide approximate inference, which is helpful to eliminate depth ambiguity. The semantic voxel encodes information about both posture and shape of clothed human, and thus provides a reasonable freedom of output space.

Integrity and detail seem to be on the opposite side in this task, the depth-guide methods avoid wrong structures but destroys details. A convincing approach deals with geometric surface details through adding a branch pipeline, to improve the robustness of recovering details. Hong et al. [HZJ*21] obtained a depth map based on the confidence volume. A relative z-offset between the sample point and its projected pixel's predicted depth is then added to this implicit function to recover details. Lahner et al. [LCT18] only exploited a two-levels, normal-based representation for generating high-frequency wrinkles detail. Our observation is that the above methods only deal with a single garment or take the clothing details as a flimsy layer, ignoring the high-quality details of multi-layer clothed human. Buffet et al. [BRB*19] proposed a method with collision-free state for untangling an arbitrary number of cloth layers, which relies on an intermediate, implicit representation. By categorizing the points as lying inside the body, between the body and the clothing, or outside the clothing, Bhatnagar et al. [BSTPM20] predicted a double-layer surface to reconstruct multi-layer clothed human. Corona et al. [CPA*21] transformed the garment template into an implicit function and tie it with the parameterized body template to finish the reconstruction for clothed human even for multi-layer case, which is a typical template-based method. Although this method infers clothing parameters to generate layered, independent garment mesh, it often does not faithfully reproduce the realistic details (such as collar, pocket) that present in the single view image because of the shortage of details module. By adding the pixel feature that are more sensitive to style details (such as collar, pocket) and additional detail feature module, our method can obtain more realistic clothing details than the template-based method [CPA*21]. Another important observation is that the existing methods [CPA*21, LIPM19, APMTM19] use the clothing semantics to divide the regions of clothing and body, but ignore the layered information contained therein, which misses the necessary part of the detailed representation. We construct a layer deformation to represent the clothed layer information of the clothed human, and recover the details of clothing on different layers.

We aim to reconstruct a multi-layer clothed human with high precision from a single view image, taking into account the changeable posture and the complex details of clothing. The single view image contains many visual information that are interlaced and interactive but deeply ambiguous. Unravelling the complex relationship of visual information and learning positive feature representation will improve the performance of 3D reconstruction. So high-precision reconstruction needs reasonable feature representation for 3D posture and cumbersome clothing details. In response to these challenges, we untangle the feature representation in the process of high-resolution clothed human reconstruction. Specifically, we use the mixed human shape features to implicitly represent the 3D pose of the clothed human. Then, the detail feature representation is constructed to restore clothing details. The highlight is that we parse the clothed layer semantics and generate the layer deformation to recover the clothing layer details. Our method can recover clothing details with geometric height at different layers and align with watertight model space. The main contributions of this work are as follows:

(1) We propose a unified feature representation for high-resolution clothed human reconstruction, which integrates the human shape feature representation and the details feature representation to finish high-quality reconstruction for clothed human with arbitrary poses and clothing details.
(2) We utilize the voxel feature with depth priors semantics to help implicit occupancy inference in the human shape feature representation, which significantly improves the robustness of 3D posture reconstruction compared with the body template.
(3) We introduce the layer deformation to represent the layer information for multi-layer clothed human and further recover high-resolution clothing details at different layers in the details feature representation, which is more accurate in the multi-layer case.

2. Related Work

2.1. RGB-D based reconstruction

Zeng et al. [ZCD*15] designed a non-rigid deformation method to compensate motion between different views, thus registering the shape of clothed human. However, registration is a strongly related problem and the dimension of naked body shape is not enough to represent the complex clothed human's shape. Another major problem is the rapid accumulation of alignment errors between two frames and the scan does not close seamlessly. An explicit "loop closure" is proposed by Wang et al. [WZD*18] to address the drifting problem in a global non-rigid registration framework. The method proposed by Zheng et al. [ZYL*18] uses a single depth camera and sparse inertial measurement units to reconstruct real-time human bodies. With this hybrid motion tracking algorithm and efficient sensor calibration technique, fast motions and challenging poses with severe occlusions can be easily recreated without the need for rigid surface reconstruction. Yu et al. [YZZ*19] and Zuo et al. [ZWZ*20] designed a two-layer structure to represent the clothed human shape, including the inner body and the outer garment mesh. The former [YZZ*19] separates the clothing template and carries out independent tracking to obtain the

body-garment independent reconstruction. The latter [ZWZ*20] uses the body parameters to guide the fusion and obtain the watertight reconstruction. Zhi et al. [ZLT*20] is a novel approach that reconstructs clothed human with full-body texture from RGB-D video. These methods achieve robust reconstruction results and provide some highly available datasets [ZYW*19, YZG*21]. However, the depth camera can work well in the experimental scene with clear background and controllable light [GLD*19, LFB17], but not in the wild. In addition, registration performs high time complexity and often fails with complex data. The RGB-D based method requires strict experimental environment and cumbersome dynamic capture and registration, the image-based reconstruction method [HLB19] has a promising prospect because of the simple input.

2.2. Parametric template representation

3D pose and shape of human have the characteristics of high dimension and strong constraint, which makes the task of body shape reconstruction learnable and easy to converge. Some methods [GRH*12, PLPM20] are proposed to generate the final reconstruction results by combining the two parts that consists of estimating the 3D pose and shape of naked body by statistical template [LMR*15, ASK*05] and simulating the wearing effect by garment templates [ZCJ*20]. Thus, Guan et al. [GRH*12] proposed an automatic 3D reconstruction method based on the parametric template. But it is limited in the garment deformation recovery caused by human pose. To tackle this problem, Patel et al. [PLPM20] predicts 3D garment deformation with three constraints including 3D pose, 3D shape, and clothing style to generate highly reliable wrinkles. Corona et al. [CPA*21] proposed a learn-based method to learn body parameters, garment parameters, and generate pose-dependent vertex deformation, which can finally combine them into the model of clothed human. In these works, garments and body are independent of each other. Thus, the ability to restoring pose-dependent deformation is constrained, and complex collision need to be handled.

There are other methods that offset the vertices of a body parametric template [LMR*15] to generate a clothed human model, but these methods are unable to simulate the complex, non-linear geometry of pose-dependent garment shapes. To address this problem, Ma et al. [MYR*20] designed a new pipeline, which learns a generative model of clothed human from 3D scans with varying pose and clothing. Multi-Garment Net [BTTPM19] is a data-driven network that completes 3D reconstruction by learning the mapping relationship between images and geometric shapes. Alldieck et al. [AMB*19] presented a learning-based model to infer the mesh of clothed human from a few frames of monocular video. The image-to-image translation methods [LIPM19, APMTM19] are novel and effective in this task which use the clothing semantics to divide the regions of clothing and body. We consider untangling the layer structure between different garment, which is challenging and the above work are not perfect. DeepHuman [ZYW*19] is an image-guided volume-to-volume convolutional neural network for 3D human reconstruction from monocular image. The above methods are sensitive to limb changes because of the pose priors. Body shape and clothing details are both recovered on the watertight surface that is limited by the number of vertices (6890) resulting in loss of many details.

2.3. Implicit representation

In the implicit method, the geometric surface is represented as a continuous energy field by inferring the occupancy probability of space. Such as PIFu [SHN*19] defines an implicit function of pixel alignment by deep learning, which transforms the regression of implicit surface into the two classifications of occupying grids, and extracts the iso-surface to generate clothed human model. As a supplement to the recovery of details, Saito et al. [SSSJ20] proposed a two-layer framework that reconstructs clothed human from high-resolution images (1024×1024) and recovers recognizable details from normal map. Huang et al. [HXL*20] is a learning-based pose perception model which finishes reconstruction from single image and applies them to animation. Furthermore, He et al. [HXS*21] introduces joint-space occupancy estimator, to improve the robustness of complex posture who sharpen depth ambiguity. He et al. [HCJS20] extracts hidden voxel and pixel features from a single image, combines them into the joint features, and calculates the probability of each space point occupied by the implicit function to estimate geometric surface. Liu et al. [LSGC21] present a framework for reconstructing 3D clothed humans from a single RGB image in which they combine an explicit and an implicit model. Xiu et al. [XYTB22] divide the details information into body details and clothing details, and reconstruct clothed human from individual video frames. The above methods do not subdivide the various features contained in the clothed human image, especially the detail module does not consider the layered information in multi-layer case, resulting in the lack of some details.

3. Method

3.1. Overview

The whole method can be divided into two parts as shown in Figure 1, the human shape feature representation and the details feature representation. Given the image of clothed human I and corresponding normal map I_n as input, the semantic image I_s and the semantic voxel V are inferred naturally by available methods [ZSZX20, LMR*15]. In voxel feature module, we estimate 3D pose from the input image and transform it into semantic voxel V, then we extract the voxel feature matrix M_v from the semantic voxel V using the 3D stacked hourglass network (3D-SHN) $E_v(\cdot)$. Similarly, the pixel feature matrix M_p is obtained from the input image I using the 2D U-Net [RFB15] $E_p(\cdot)$ in the pixel feature module. In details feature module, we parse the clothing semantic from the superposition of input image I and semantic image I_s using VGG16, and build the layer deformation that is responsible for providing the clothed layer feature. The normal feature is extracted from the normal image I_n using the 2D stacked hourglass network (2D-SHN) $E_d(\cdot)$. A set of features is obtained for each query point P. The voxel feature and the pixel feature are aligned to the human shape feature representation. The clothed layer feature and the normal feature are aligned to the details feature representation. Multi-layer perceptron (MLP) with parameters θ is used to implement the implicit surface function $f_\theta(\cdot)$. A continuous function $F(P) \in [0, 1]$ to deduce the occupancy of query point P with the human shape feature representation and the details feature representation. It is formulated in Equation (1). Finally, the geometrical reconstruction

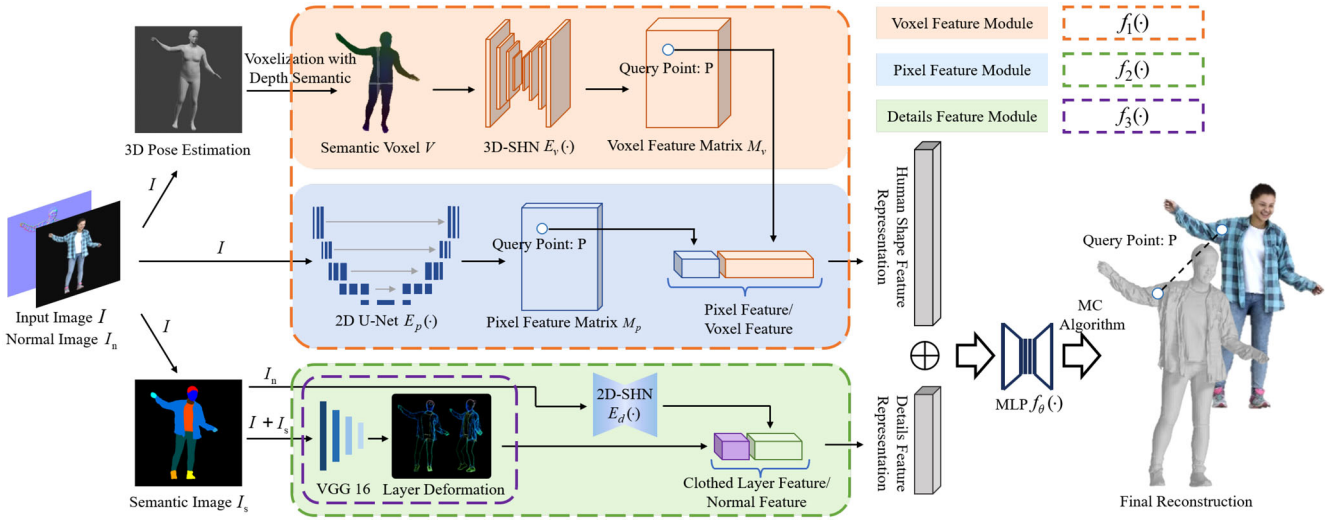

Figure 1: *The pipeline of proposed method. We combine the pixel feature and the voxel feature extracted from semantic voxel to represent human shape. The details feature module constructs the layer deformation from the inferred clothed layer semantics to obtain the clothed layer feature. The clothing details are restored by using the details feature mixed with the clothed layer feature and the normal feature on different layers.*

result can be drawn from implicit surfaces by Marching Cubes (MC) algorithm [LC87].

$$f_\theta(f_1(E_v(V), E_p(I), P), f_2(f_3(I, I_s), E_d(I_n), P)) \mapsto F(P) \quad (1)$$

3.2. Human shape feature representation

The shape of clothed human is limited by arbitrary posture and clothing, and the reconstruction for clothed human can not be completed only by inferring 3D posture and body shape parameters. We propose the human shape feature representation $f_1(E_v(V), E_p(I), P)$ in Equation (1) to integrate the latent code of human shape, which is composed of the voxel features and the pixel features respectively. For each query point P, voxel feature provides shape and depth information to constrain the degrees of freedom of limb and invisible back shape. As a supplement, pixel feature provides fine-grained geometric information, which helps to generate geometric surfaces with clothing details rather than smooth skin.

3.2.1. *Voxel feature*

$E_v(V)$ in the human shape feature representation encodes the semantic voxel V into a voxel feature matrix M_v. M_v is aligned with the model space of the final output result, so we can separate it by marching cube algorithm to get a rough result. During inference, we store the latent code and depth value as voxel features. Compared with the existing methods [XYTB22, ZYLD21], our semantic voxel has three advantages: (1) it encodes inherent position about both the shape and pose of the clothed human, thus provides a reasonable freedom of output space, (2) the semantic provides approximate depth to help voxel feature inference and (3) it is easy to obtain from the image and does not depend on the accuracy of the datasets mesh.

Specifically, we firstly exploit the method proposed by Zhu et al. [ZZW*19] to estimate 3D pose from input image I. Then, we render the depth semantic code onto the image plane to obtain a semantic map and generate a semantic voxel by voxelization of SMPL [LMR*15]. Finally, we propagate the semantic codes into the occupied voxel. The 3D-SHN is composed of four repeated standard hourglasses. Each hourglass obtains the size invariant tensor and the heat map as the next input and intermediate supervision, respectively. We simplify the cross-entropy error (CEE) in He et al. [HCJS20] and take 6, 890 vertices on the surface S of SMPL as sampling points in training.

$$CEE = \frac{1}{n}\sum_{P\in S} f(P_i^*)\cdot log(f(P_i)) + (1 - f(P_i^*))\cdot log(1 - f(P_i)) \quad (2)$$

where n is the number of point samples, P_i is a 3D point sample indexed by i, P_i^* is a ground truth point indexed by i, $f(\cdot)$ computes the predicted occupancy value for query point P or P^*.

3.2.2. *Pixel feature*

$E_p(I)$ in the human shape feature representation encodes the input image I into the latent space and obtains the pixel features of the query point P. The pixel-aligned implicit function was first proposed by PIFu [SHN*19] and widely used in similar methods [SSSJ20, HZJ*21]. These methods usually require an encoder with a wide receptive field in order to support overall perception and consistent inference of depth. However, the large receptive field cannot pay attention to local, fine feature, contrary to expectation that generates high-fidelity result. In this grid, we store information about surfaces of a small local neighbourhood in the form of independent latent codes. In addition to simplifying the prior distribution that the network must learn, this decomposition of scenes into local shapes also

makes inference more efficient. In the training stage, we encode the input image I into the pixel feature matrix M_p by 2D U-Net [RFB15] with the cross-entropy error [HCJS20]. For each query point P, M_p provides the local depth features and depth values from local volume as pixel feature, instead of global depth feature like PIFu [SHN*19].

3.3. Details feature representation

The surface details of clothed human are distributed on different clothing layers, which brings difficulties to the representation for geometric details [CPA*21]. To represent the varying clothing information on different layers, we propose the details feature representation $f_2(f_3(I, I_s), E_d(I_n), P)$ in Equation (1) which is two levels representation to approach toward high-resolution result for clothed human. Specifically, $f_3(\cdot)$ uses the layer deformation to implicitly represent the clothed layer information of clothed human, and compute the deformation from different semantics to predict the clothed layer feature on geometric surface. Then, $E_d(\cdot)$ encodes the normal feature as the proxy for 3D geometry. Finally, the reconstruction is guided by the details feature representation which mixes with the clothed layer feature and the normal feature to predict a 3D geometry, making it easier for the MLP to produce details.

3.3.1. *Layer deformation*

In daily life, people wear clothes in a certain order. For top, T-shirt is on the inner layer, coat is on the outer layer, and sweater is in the middle layer. For bottom, leggings are inside and dress is outside [YLL14]. This observation is a more restrictive attribute than the clothing semantics. We analyse the clothing semantics labels in the datasets proposed by Yang et al. [YLL14] that contain 58 complex labels, including top, bottom, shoes and accessories. We delete the labels of shoes and divide the rest into three categories: top (t), bottom (b), and accessories (a). The top and bottom contain three layers, as shown in Table 1. The layer semantics corresponding to the same clothing semantics is uncertain (e.g. the coat may be worn in the second or third layer. Dress or shorts is the second layer when wearing tights.). The clothing semantic $s_x \in S_{clothing}$ for each pixel x can be transform to S_{layer}. These prioris can be formulated as follows:

$$S_{layer} = \{(x, f(s_x))\} \tag{3}$$

In Equation (3), $f(s_x)$ denotes the transformation of semantic information, such as $f(\text{T-shirt}) < f(\text{sweater}) < f(\text{coat})$, which is inferred from the above prioris. Based on these rules and the semantic labels contained in Table 1, we define the constraint knowledge φ of daily clothing as follows.

$$\varphi = \{t_1, t_2 \cdots t_{21}\} \cup \{b_1, b_2 \cdots b_{11}\} \cup \{a_1, a_2 \cdots a_{15}\} \tag{4}$$

For top t and bottom b, φ contains the transformation rules of different layer. Such as we infer $t_1 > t_3$, which means t_1 is the upper layer than t_3, from $t_1 > t_2$ and $t_2 > t_3$. For accessories a, φ contains the relation of dependence. Such as we infer $a_1 > t_1$, which means a_1 is the upper layer than t_1, from $a_1 \in t_1$.

Ideally, φ infers the layer labels from the parsing result that contains rich semantics. However, it is difficult for the existing datasets to cover all clothing styles, and complex labels can also fragment regions. To solve the problem of having difficulties in understanding the multiple layers of unlabelled or complex semantics, we propose a segmentation method for clothed human that can automatically predict the layer information of clothing. Specifically, we take the joint points as the landmark of different regions, and sample the clothing region. Then, we compute the average of Euclidean distance $\|\|^2$ between the coordinate of sample point p_i^k and the coordinate of landmark m in each semantic area k, as the evidence of predicting. This process can be formulated as follows:

$$E_k = avg\left\| p_i^k - m \right\|^2 \tag{5}$$

$f_3(I, I_s)$ in Equation (1) consists of two main steps: (1) Parsing the clothed layer semantics. (2) Calculating deformation offset. Following Zhang et al. [ZSZX20] work, the clothing semantic image I_s can be predicted and the joint points can be extracted by OpenPose [CSWS17] from input image I. Then we use VGG16 network to parse the clothed layer semantics $S_{clothing}$ from the superposition of clothing semantic images I_s and clothed human images I. In order to complete the transformation from the clothing semantics $S_{clothing}$ to the clothed layer semantics S_{layer}, we use joint points to calculate the Euclidean distance of each clothing semantics E_k, and take it as the optimization item of the loss function. Specifically, the midpoint of neck and hip joint is used as the landmark of top area, wrist joint as the landmark of sleeve area, and ankle joint as the landmark of bottom area. The layer of semantic is higher with the increase of E_k. More details see Figure 2.

We train VGG16 network with the mixed cross entropy error (Equation 6) that calculates the cross entropy between the ground truth y_i^* and the predicted clothing semantic y_i to predict the clothed layer semantic S_{layer}. The performance is improved by minimizing the cost function $\lambda \sum_k^N E_k$ to ensure reasonable results in training. Different from the existing method [YLL14] that parses wide range of clothing semantics and contains a large number of labels (Table 1), our proposed method only classifies $N = 8$ regions of layer, including background, skin, accessories, first-top,

Table 1: *Clothing semantic labels and clothed layer semantic labels*

	Top (21)	Bottom (11)	Accessories (15)
First layer (31)	Intimate Bra T-shirt Shirt Blouse Vest Sweatshirt Swimwear	Shorts Pants Dress Tights Romper Jeans Leggings Panties	Wallet Tie Cape Watch Ring Scarf Gloves Hat Belt Earrings Glasses Necklace Bracelet Purse Sunglasses
Second layer (13)	Blazer Coat Hoodie Cape Jumper Jacket Sweater Suit Bodysuit Cardigan	Dress Shorts Bodysuit	None
Third layer (3)	Coat Cape Suit	None	None

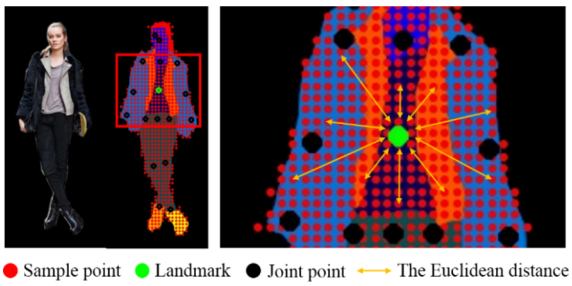

Figure 2: *We explain the proposed method on an enlarged version of the semantic diagram. Calculating the average of the Euclidean distance between the sampling point and the landmark in each region, we can infer the clothed layer semantic. The distance of the uppermost layer (blue) is the largest, and the distance of the lowermost layer (purple) is the smallest.*

second-top, three-top, first-bottom, second-bottom, which is easy to learn through conventional segmentation algorithm.

$$LOSS = \sum_i y_j^* \cdot log(y_i) + \lambda \sum_k^N E_k \tag{6}$$

Based on the clothed layer semantic S_{layer}, we construct the layer deformation to represent clothed layer information of clothed human. Essentially, the layer deformation is a vector-valued function that represents a latent vector field. Each latent vector β on geometric meshes M_k consists of the normal direction of the original vertex $v_i \in Ver$ and the deformation increment that forms height difference on geometric surface. Unlike Huang et al. [HXL*20] that predict the global displacement of vertices with joint, pose and skinning weights, our method infers the deformation increment with different semantics $s_i \in S_{layer}$ by computing the deformation $T(s_i)$. Then, the deformation propagates along the normal direction of the original vertex v_i. The vector-valued function define as:

$$f(Ver, S_{layer}) = \sum_{i=1} T(s_i)v_i \tag{7}$$

3.3.2. *Normal feature on different layers*

The clothing details of the multi-layer clothed human are distributed at different layers, which is difficult to untangle and recover them. We take the latent vector β representing the height difference in Section 3.3.1 as the clothed layer feature to recover the clothed layer details, and further use $E_d(\cdot)$ encodes the normal feature from normal map I_n to represent the folds at different layers. As a supplement to the human shape feature, the predicted details feature for each query point P is composed of the clothed layer feature and the normal feature, which helps to infer the geometric surface with style, clothed layer and wrinkles.

The stacked hourglass network is used as the encoder $E_d(\cdot)$. The residual structure of hourglass module can effectively extract multi-scale features that extend the image channel into the stereo depth space. It is a size invariant transformation network that ensures the size of details feature is aligned with the output space. The mean square error (MSE) is used during training and densely sample on the ground truth mesh by the geometric sampling strategy.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} |F(P_i) - F(P_i^*)|^2 \tag{8}$$

where n is the number of samples, P_i is a 3D sample indexed by i, P_i^* is a ground truth point indexed by i, $F(\cdot)$ computes the occupancy value of query point P_i or P_i^*.

4. Experimental Results and Analysis

We complete our experiment on a NVIDIA GeForce RTX 2080ti GPU with 32GB DDR4 2666MHz RAM and the system is Ubuntu 16.04. The convolution neural network is constructed by PyTorch 13.0.1.

4.1. Datasets

Different from image to image vision algorithm, 3D reconstruction algorithm relies on high-resolution 3D model as ground truth supervision [FAZ21]. The existing 3D datasets of clothed human are divided into two categories: the scanned watertight models (e.g. RenderPeople, TWINDOM, Yu et al. [YZG*21]) and the synthetic models (e.g. 3DPeople [PSRC*19], CLOTH3D [BME20]).

In order to provide high-precision 3D supervision from ground truth mesh, we use the datasets built by us and the datasets proposed by Yu et al. [YZG*21] as training datasets. In addition, we use the datasets proposed by Zhang et al. [ZPBPM17] as a supplementary evaluation, which is a completely unlearned 3D people in training stage. Furthermore, we select web-images to verify the proposed method, which can qualitatively evaluate the robustness of our method to posture and clothing details. Considering that the low-resolution images lack information and the high-resolution images cause memory constraints, we resize all images to 512×512 size as input.

4.2. Comparison and analysis

4.2.1. *Quantitative analysis*

We use four popular evaluation metrics provided by He et al. [HCJS20] to quantitatively compare with DeepHuman [ZYW*19], PIFu [SHN*19], and Saito et al. [SSSJ20] on different datasets [YZG*21, ZPBPM17]. During evaluation, all methods use the pre-trained model provided by the author. We use the iterative nearest point algorithm to register the results with the ground truth (GT). Then, we measure the metrics follows:

- **P2S**: The average point-to-surface Euclidean distances (signed/unsigned) from the vertices on the reconstructed surface to the GT.
- **L2**: The normal L2 distances to evaluate the reconstruction accuracy.
- **CD**: The Chamfer distances between the result and the GT surfaces to evaluate minimum reconstruction error.
- **HD**: The Hausdorff distances between the result and the GT surfaces to evaluate maximum reconstruction error.

Table 2: *Quantitative evaluation of related methods on different datasets*

	Our datasets				Yu et al. [YZG*21]				Zhang et al. [ZPBPM17]			
	P2S*	L2	CD	HD	P2S*	L2	CD	HD	P2S*	L2	CD	HD
DeepHuman [ZYW*19]	8.80	-	4.50	9.55	8.57	-	4.44	9.30	5.33	-	3.30	8.23
PIFu [SHN*19]	4.43	0.34	3.02	8.85	4.23	0.33	2.98	8.80	1.15	0.09	1.14	8.65
PIFuHD [SSSJ20]	4.21	0.30	2.98	8.75	4.01	0.30	2.73	8.74	1.63	0.13	1.75	7.60
Ours (Human shape)	4.20	0.29	3.01	8.55	4.15	0.32	2.74	8.53	1.50	0.18	1.30	7.44
Ours (Human shape+Details)	4.20	0.25	2.99	8.55	4.10	0.29	2.73	8.53	1.47	0.13	1.31	7.44

*It is an unsigned P2S.

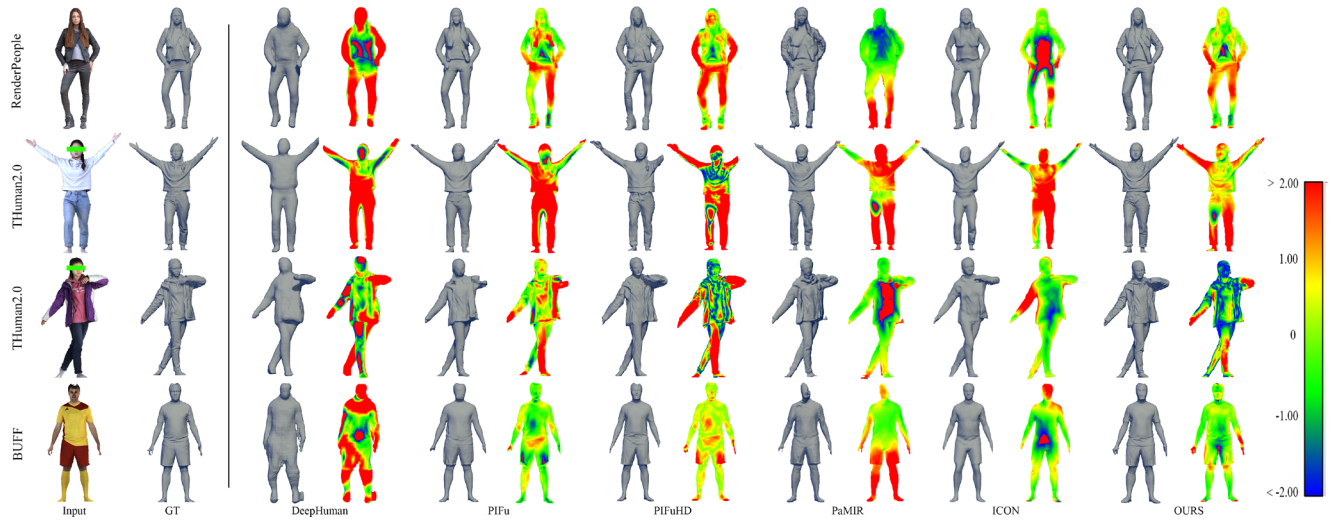

Figure 3: *Qualitative comparison results of our method and other methods [ZYW*19, SHN*19, SSSJ20, ZYLD21, XYTB22]. Columns 4, 6, 8 and 10 show the visualization of signed P2S error. The maximum threshold is set 2, which means that the distance error greater than 2 or less than -2 cannot be distinguished by colour. The experimental results show that the proposed method has higher fidelity than the related methods.*

The results shown in Table 2. Compared with the voxel-based method like DeepHuman [ZYW*19], proposed method shows superiority in four metrics. Compared with the image-based method [SHN*19, SSSJ20], our method achieves significant reduction in HD metric. This error comes from the wrong structure or fragmentation of limbs, and our method can effectively solve this problem. Compared with PIFu [SHN*19] on our datasets and the datasets proposed by Yu et al. [YZG*21], the HD error of our method is reduced by 3.16% and 3.07%, respectively. L2 and CD are used to evaluate the accuracy of reconstruction, in these parts, our method is similar to Saito et al. [SSSJ20].

We completed the signed P2S error analysis of our method and the state-of-the-art methods on cloud compare,[1] then drew the visualization results (as shown in Figure 3). DeepHuman [ZYW*19] uses single voxel feature to infer the human shape, ignoring the representation of the clothing visual information, and only reconstructs the coarse posture without surface details. Furthermore, the expansion of voxels is negative for the extremities and surface details. PIFu [SHN*19] uses single pixel feature to infer the human shape, which is perspicacious of surface details but unable to bridge the gap of deep ambiguity and to adapt to changes in limbs. Saito et al. [SSSJ20] uses the features extracted from high-resolution images (1024×1024) to guide fine-grained reconstruction, and uses the predicted front and back normal maps to enhance the performance of detail recovery. In these pixel-based methods [SHN*19, SSSJ20], due to the lack of effective three-dimensional constraints, many errors appeared on the invisible back and limbs. [ZYLD21] and [XYTB22] are outstanding contributions of the hybrid feature based reconstruction. After the fusion of voxel features, the robustness of the reconstruction to pose has been significantly improved. As a result, proposed method is more accurate than either voxel-based or pixel-based methods. In summary, our method achieves a high fidelity digital clothed human.

[1] https://www.cloudcompare.org

4.2.2. *Qualitative analysis*

We select representative from the related works and compare them with our method. Figure 4 demonstrates the comparison results with the fusion-based method [ZWZ*20]. The experimental results show that our method restores more details on some simple imitation examples. However, some errors are highlighted that the texture is incorrectly identified as wrinkles. We also exhibit the results of related works and ours, see Figure 5. The voxel

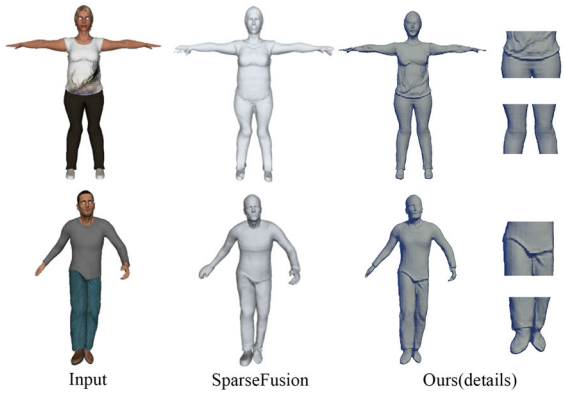

Figure 4: *Comparison with the fusion-based method [ZWZ*20]. Our method achieves better detail restoration, including layered detail and wrinkles. Some errors are highlighted due to texture interference (woman's T-shirt).*

provides rough human shape information, which is positive for body structure, but blurs the surface clothing details, as shown in the results of DeepHuman [ZYW*19]. The template-based method uses parameterized template [LMR*15] to estimate posture and restores clothing details by offsetting vertices [CPA*21]. These methods can not reconstruct realistic human model from a single image. The pixel-based methods [SHN*19, SSSJ20] infer the implicit expression of clothed human shape and achieve advanced detail recovery. The hybrid-based methods [ZYLD21, XYTB22] have excellent performance, but there are still some unpredictable error representation. In contrast, our method reconstructs complete posture and limb ends, and recovers the multi-layer detail with height difference, which is better than the related methods.

4.3. Ablation study and analysis

We highlight the improvement of the feature representation by using different MLP and the unsigned P2S to evaluate the results (shown in Table 3). MLP01's size is [257, 1024, 512, 256, 128, 1], MLP02's size is [257, 1024, 1024, 512, 256, 128, 1], MLP03's size is [257, 512, 1024, 1024, 512, 256, 128, 1]. The experimental results show that the feature representation proposed in our paper is effective for high-resolution clothed human reconstruction. According to the experimental results and computational power constraints, we choose MLP01 as $f_\theta(\cdot)$ in Equation (1).

As shown in Figure 6, benefiting from the robustness of voxel priors to 3D pose estimation, our method can reconstruct a 3D clothed human with challenging pose rather than simple A or T pose. The depth blurring between the limbs is difficult to be restored by a single pixel feature [SHN*19]. In addition, the fine-grained pixel feature and the details feature help to infer geometric surface details, which is important for clothed human reconstruction. Figure 7 shows the applicability of our method for clothing details and small accessories and our method can finish high-resolution reconstruction for clothed human.

Table 3: *Effectiveness evaluation of the feature representation using different MLP*

	Our datasets	Yu et al. [YZG*21]	Zhang et al. [ZPBPM17]
MLP01	4.20	4.10	1.47
MLP02	4.20	4.20	1.48
MLP03	3.99	4.20	1.48

To demonstrate the importance of human shape feature representations and details feature representations for high-fidelity inference, we conduct an ablation study (as shown in Figure 8). We use a simple trained MLP to infer geometric results from the voxel feature and the human shape feature, which contain human shape (the 2–5 columns in Figure 8). The last two columns in Figure 8 provide the final result of our method. The single voxel feature can constrain the reconstruction result to a reasonable degree of freedom, but it can only deal with the approximate human shape without clothing details. Due to the expansion of voxels, the visualization results look bloated. Inversely, the pixel feature constrains the shape within the query contour range. The mixed human shape feature shows excellent inference performance in implicit expression. Furthermore, the experiment results show that the detail feature optimization designed by us is effective for clothed human, and can restore high-precision surface details scattered at different layers. The voxel features provide priors to help deal with challenging poses, such as Figure 8a and b. The mixed human shape feature and details feature optimization enhance details representation which can restore style details and small accessories, such as hat, belt in Figure. 8c. Our method is also good at rich wrinkles scene such as Figure 8d.

As shown in Figure 9, we use web-images to evaluate the effectiveness of our method. The model in these images is dressed in complex and without geometric model as GT. We learn simple layer semantic labels to replace complex clothing semantic labels that can demarcate accurate areas, but the connotation is not accurate. For example, long skirt (Figure 9 c) and pleated skirt (Figure 9b) are all considered dress. On the contrary, our method cares about the layer rather than the type of clothing. Our method needs the priori method [ZSZX20] to provide the semantic integrity of clothing rather than fragmentation. For the parsing results of new clothing semantics that have never appeared, the adaptability of our method is poor. Furthermore, the detailed feature module computes offsets of the different layers and restores clothing details with geometric height after learning the layer semantics. We also show the application prospect of real digital human, more results of the proposed method are shown in Figure 10, including the display of different perspectives of reconstruction results and the full-body textured results. Our method can reconstruct 3D clothed human with accurate posture and rich details, which shows potential in the field of 3D virtual try-on and digital character.

5. Conclusions

The performance of clothed human reconstruction is limited by the changes in posture and the intricate variations in clothing. In order to solve these two problems comprehensively, a unified feature representation for clothed human is proposed in this paper, which integrates shape and details feature representation to reconstruct clothed human. Specifically, the human shape feature

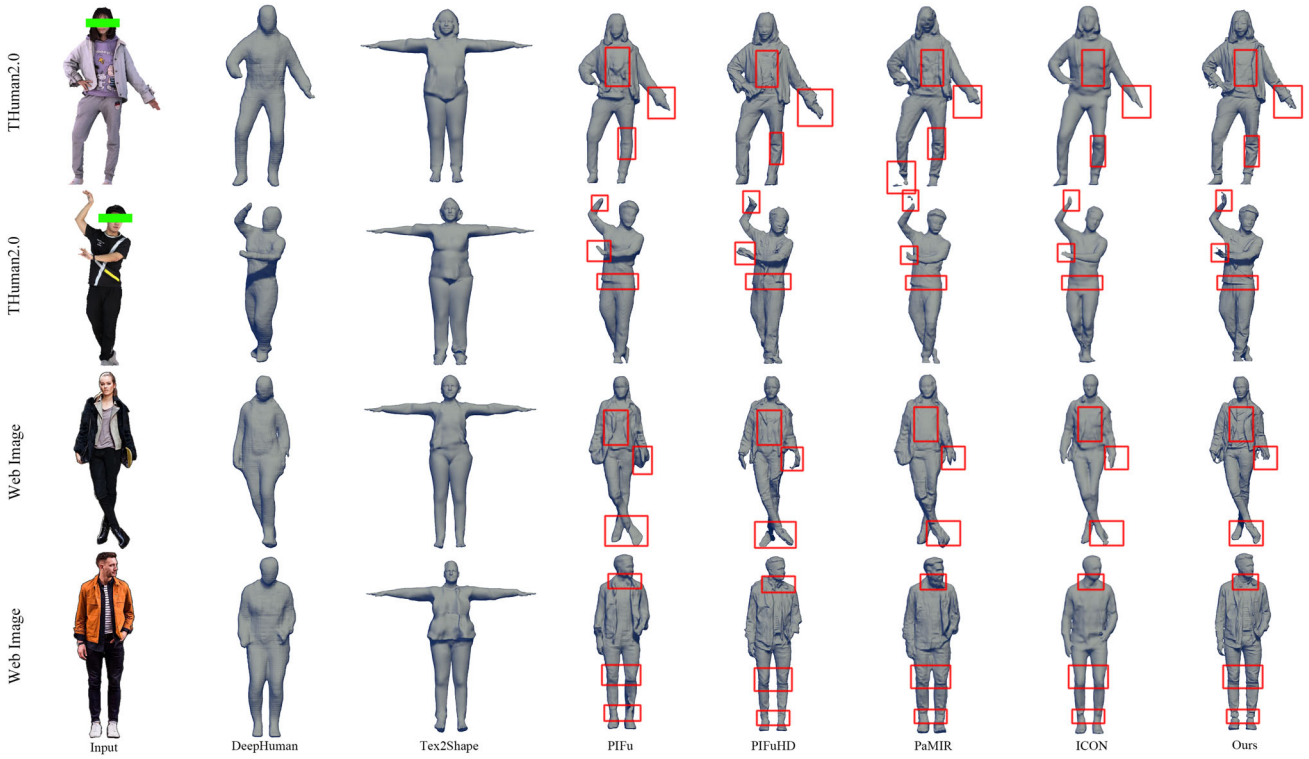

Figure 5: *Qualitative comparison with other methods [ZYW*19, APMTM19, SHN*19, SSSJ20, ZYLD21, XYTB22]. Our method credibly reconstructs the geometric mesh for clothed human with limbs and clothing details (red box).*

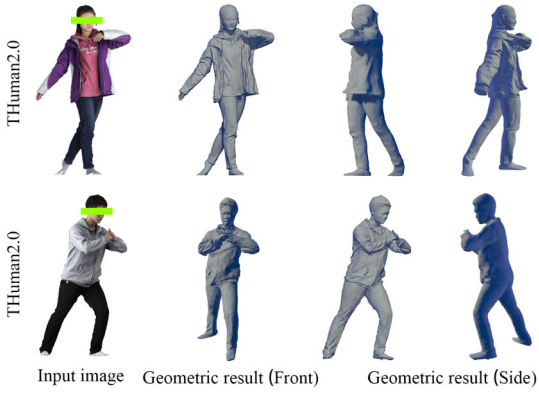

Figure 6: *We utilize the voxel features with depth priors semantics to help implicit occupancy inference, which significantly improves the robustness of 3D posture reconstruction for clothed human. Our method can reconstruct a geometric result with challenging posture.*

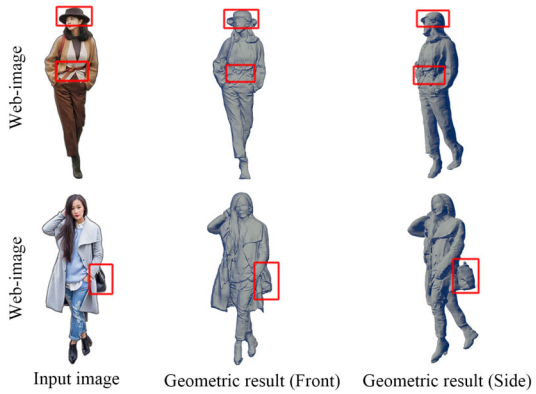

Figure 7: *Our method can reconstruct high-precision geometric details, including clothing details and some small accessories (such as hat, belt, handbag).*

representation uses the semantic voxel feature to constrain the inference of body shape within a reasonable degree of freedom, which can improve the accuracy of reconstruction for limb changing and invisible back. Moreover, the details feature representation untangle the layered details and restore high-resolution clothing details at different layers, which promotes the accuracy and fidelity. Our method can cope with changeable posture and complex clothing details, but it still fails in the case of expansively challenging posture (as shown in Figure 11a) or unpredictable large clothing and accessories (as shown in Figure 11b and c). This is because extreme posture, large clothing and accessories tear the depth gap between two to three-dimensional representation and destroy the latent dimensional space of results. Furthermore, we infer three layers clothing semantic that is rely on prioris accuracy. For cases with more than three layers, our method is powerless. Our method relies on the absolute correctness

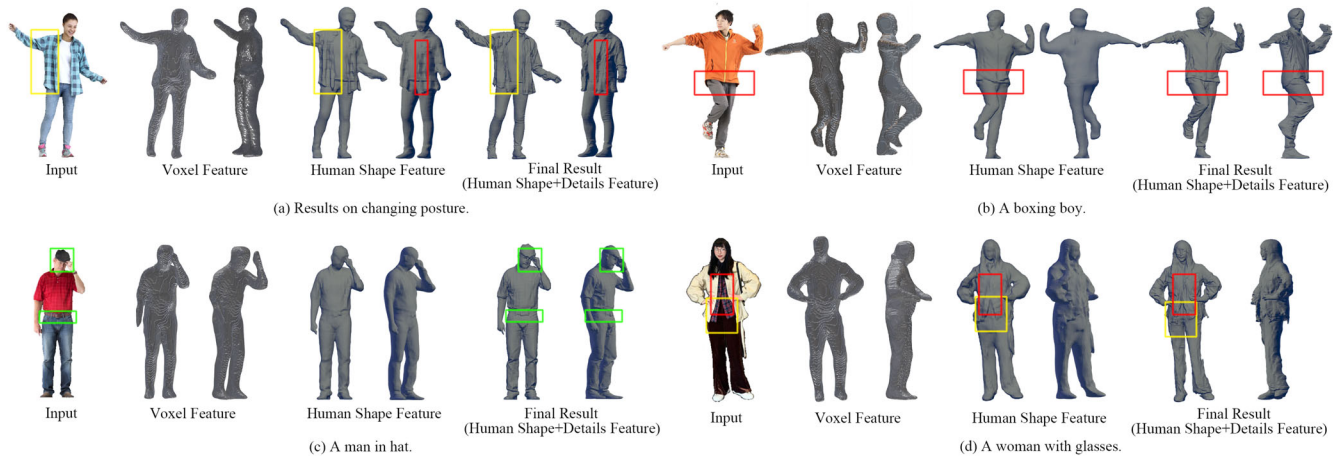

Figure 8: *Ablation study on our method. The inference performance of the human shape feature mixed with voxel and pixel feature is better than the single voxel feature's. Our method can restore style details and small accessories (green box). The details feature representation contributes to high-precision inference which reduces the collapse caused by textured mapping (yellow box) and is adaptable for the multi-layer case (red box). We manually adjust the joint points to reduce the large errors in the attitude prior, see case (c).*

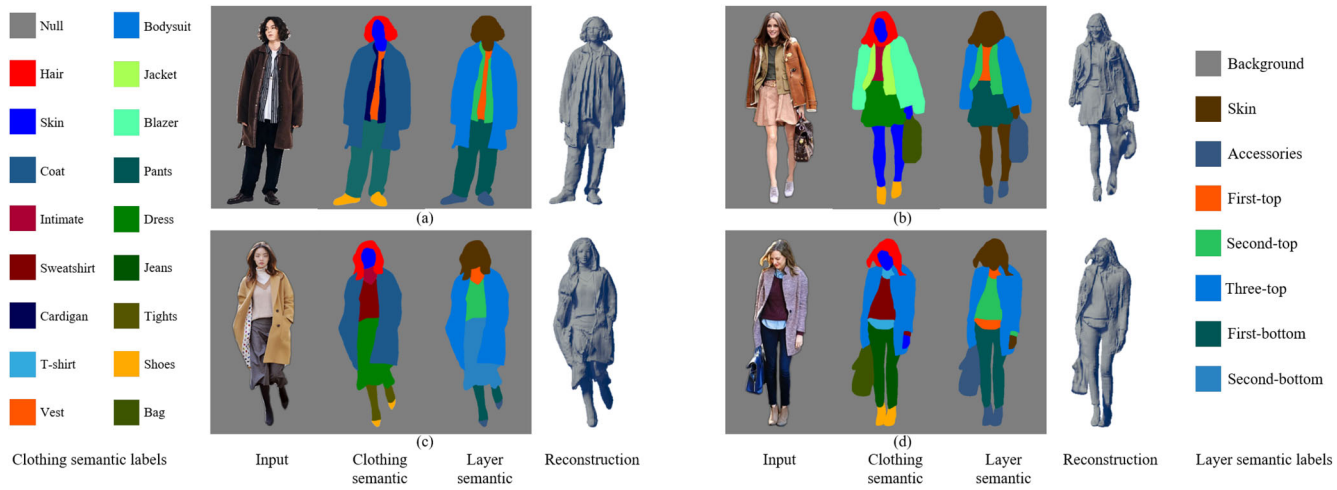

Figure 9: *We simplify the clothing semantic labels (left) and use eight layer semantic labels (right) to indicate the order of dressing. The detail feature module can restore clothing details at different layers and finally generate a hierarchical clothed human model.*

of clothing semantic parsing and 3D posture estimation rather than improving them, and their errors will spread to our method. Reconstruction relies on high-precision 3D datasets that can not contain all poses and clothing, which challenges the generalization of clothed human model. Besides, we find that the reconstruction results obtained from render images are better than those from real scenes. These input images look similar, get different results. A reasonable explanation is that the render image is the calculation product from geometric model, texture and controllable lighting, while the real image is the induction product of optical elements and the distortion caused by perspective affects the reconstruction (as shown in Figure 11d). In the design of detail module, how to untangle clothing texture and wrinkles, to avoid misjudgment is still a problem to be solved (as shown in Figure 4, woman's T-shirt. Figure 5, second line, boy's T-shirt). This work lacks the adaptive performance of seasonal clothing.

An important direction of future work is to get rid of the dependence on high-precision datasets and improve the robustness of reconstruction method for real images by exploring 3D inference under weak supervised or unsupervised environment. These are also our goals that remove artefacts in voxel representation to get more suitable body shape feature and improve the detail restoration accuracy of clothing for special occasions (fashionable evening dresses, ethnic minority clothing, etc.) and different season's clothing. The former can be solved by obtaining more accurate body weight from the image, and the detailed feature representation of a specific scene can be considered to deal with the latter. Moreover, our method

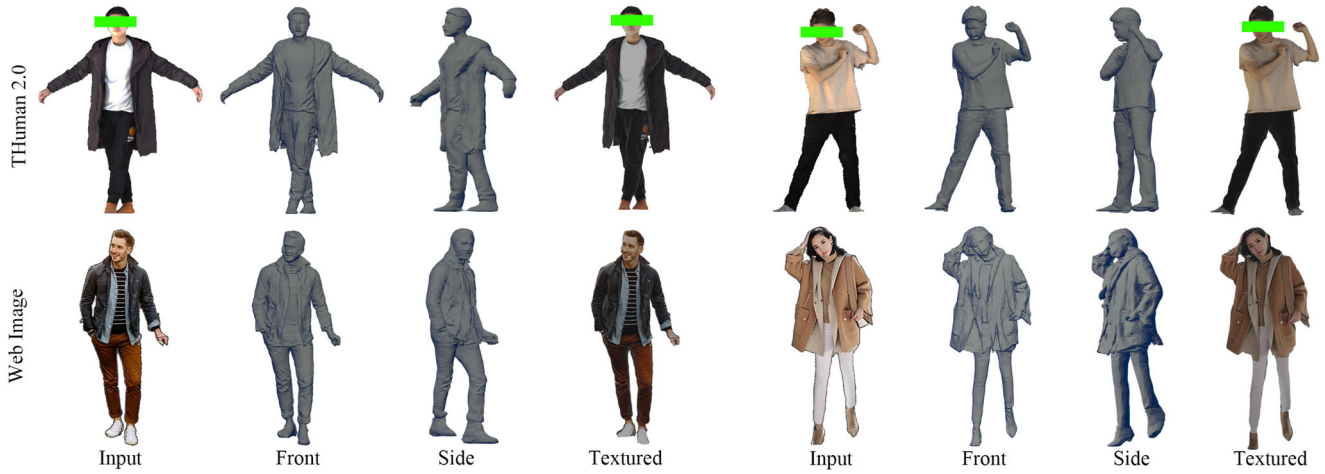

Figure 10: *We show the reconstruction results from different perspectives. The last column, we provide the results after stitching the texture map in blender. Our approach shows promising potential in fields such as digital character generation.*

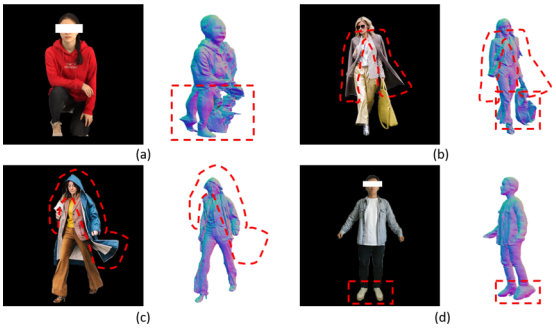

Figure 11: *Fails case. Our method fails in the case of extremely challenging posture or unpredictable large accessories. The distortion caused by perspective also affects the accuracy.*

focuses on posture and clothing details, but has errors in hair, head and other details which are worth considering in future work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62262036, 61862036, 61962030) and the Yunnan Provincial Foundation for Leaders of Disciplines in Science and Technology (202005AC160036).

References

[AMB*19] 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), pp. 1175–1186.

[APMTM19] 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), pp. 2293–2303.

[ASK*05] Anguelov D., Srinivasan P., Koller D., Thrun S., Rodgers J., Davis J.: Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*. (2005), pp. 408–416.

[BME20] Bertiche H., Madadi M., Escalera S.: Cloth3d: clothed 3d humans. In *European Conference on Computer Vision*. Springer, (2020), pp. 344–359.

[BRB*19] Buffet T., Rohmer D., Barthe L., Boissieux L., Cani M.-P.: Implicit untangling: A robust solution for modeling layered clothing. *ACM Transactions on Graphics (TOG) 38*, 4 (2019), 1–12.

[BSTPM20] Bhatnagar B. L., Sminchisescu C., Theobalt C., Pons-Moll G.: Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*. Springer, (2020), pp. 311–329.

[BTTPM19] 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 Ccomputer Vision* (2019), pp. 5420–5430.

[CPA*21] 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), pp. 11875–11885.

[CSWS17] Cao Z., Simon T., Wei S.-E., Sheikh Y.: Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 7291–7299.

[FAZ21] Fahim G., Amin K., Zarif S.: Single-view 3d reconstruction: A survey of deep learning methods. *Computers & Graphics 94* (2021), 164–190.

[GLD*19] Guo K., Lincoln P., Davidson P., Busch J., Yu X., Whalen M., Harvey G., Orts-Escolano S., Pandey R., Dourgarian J., et al.: The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG) 38*, 6 (2019), 1–19.

[GRH*12] Guan P., Reiss L., Hirshberg D. A., Weiss A., Black M. J.: Drape: Dressing any person. *ACM Transactions on Graphics (ToG) 31*, 4 (2012), 1–10.

[HCJS20] He T., Collomosse J., Jin H., Soatto S.: Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems 33* (2020), 9276–9287.

[HLB19] Han X.-F., Laga H., Bennamoun M.: Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 5 (2019), 1578–1604.

[HXL*20] 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), pp. 3093–3102.

[HXS*21] He T., Xu Y., Saito S., Soatto S., Tung T.: Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 11046–11056.

[HZJ*21] 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), pp. 535–545.

[LC87] Lorensen W. E., Cline H. E.: Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics 21*, 4 (1987), 163–169.

[LCT18] Lahner Z., Cremers D., Tung T.: Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 667–684.

[LFB17] Leroy V., Franco J.-S., Boyer E.: Multi-view dynamic shape refinement using local temporal integration. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 3094–3103.

[LIPM19] Lazova V., Insafutdinov E., Pons-Moll G.: 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*. IEEE, (2019), pp. 643–653.

[LMR*15] 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*, 6 (2015), 1–16.

[LSGC21] Liu L., Sun J., Gao Y., Chen J.: Hei-human: A hybrid explicit and implicit method for single-view 3d clothed human reconstruction. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, (2021), pp. 251–262.

[MYR*20] 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), pp. 6469–6478.

[PLPM20] 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), pp. 7365–7375.

[PSRC*19] Pumarola A., Sanchez-Riera J., Choi G., Sanfeliu A., Moreno-Noguer F.: 3dpeople: Modeling the geometry of dressed humans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 2242–2251.

[RFB15] Ronneberger O., Fischer P., Brox T.: U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention* (2015), pp. 234–241.

[SHN*19] 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 (ICCV)* (2019), pp. 2304–2314.

[SSSJ20] 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 (CVPR)* (2020), pp. 84–93.

[WZD*18] Wang S., Zuo X., Du C., Wang R., Zheng J., Yang R.: Dynamic non-rigid objects reconstruction with a single RGB-D sensor. *Sensors 18*, 3 (2018), 886.

[XYTB22] Xiu Y., Yang J., Tzionas D., Black M. J.: Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 13296–13306.

[YLL14] Yang W., Luo P., Lin L.: Clothing co-parsing by joint image segmentation and labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 3182–3189.

[YZG*21] Yu T., Zheng Z., Guo K., Liu P., Dai Q., Liu Y.: Function4d: Real-time human volumetric capture from very sparse consumer RGBD sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 5746–5756.

[YZZ*19] Yu T., Zheng Z., Zhong Y., Zhao J., Dai Q., Pons-Moll G., Liu Y.: Simulcap: Single-view human performance

capture with cloth simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Eecognition* (2019), pp. 5504–5514.

[ZCD*15] Zeng M., Cao L., Dong H., Lin K., Wang M., Tong J.: Estimation of human body shape and cloth field in front of a kinect. *Neurocomputing 151* (2015), 626–631.

[ZCJ*20] 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), pp. 512–530.

[ZLT*20] Zhi T., Lassner C., Tung T., Stoll C., Narasimhan S. G., Vo M.: Texmesh: Reconstructing detailed human texture and geometry from RGB-D video. In *European Conference on Computer Vision*. Springer, (2020), pp. 492–509.

[ZPBPM17] Zhang C., Pujades S., Black M. J., Pons-Moll G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4191–4200.

[ZSZX20] Zhang Z., Su C., Zheng L., Xie X.: Correlating edge, pose with parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8900–8909.

[ZWG*21] Zhang Q., Wang L., Ge L., Luo S., Zhu T., Jiang F., Ding J., Feng J.: A robust multi-view system for high-fidelity human body shape reconstruction. *Computer Graphics Forum: Journal of the European Association for Computer Graphics 40*, 5 (2021), 19–31.

[ZWZ*20] Zuo X., Wang S., Zheng J., Yu W., Gong M., Yang R., Cheng L.: Sparsefusion: Dynamic human avatar modeling from sparse RGBD images. *IEEE Transactions on Multimedia 23* (2020), 1617–1629.

[ZYL*18] Zheng Z., Yu T., Li H., Guo K., Dai Q., Fang L., Liu Y.: Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 384–400.

[ZYLD21] 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 44*, 6 (2021), 3170–3184.

[ZYW*19] 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), pp. 7739–7749.

[ZZW*19] Zhu H., Zuo X., Wang S., Cao X., Yang R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 4491–4500.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information