VS: Reconstructing Clothed 3D Human from Single Image via Vertex Shift

Leyuan Liu[1], Yuhan Li[1], Yunqi Gao[1], Changxin Gao[2], Yuanyuan Liu[3], Jingying Chen[1,*]
[1] National Engineering Research Center for E-Learning, Central China Normal University
[2] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[3] School of Computer Science, China University of Geosciences (Wuhan)
{lyliu,chenjy}@mail.ccnu.edu.cn, {lyh985812,gaoyunqi}@mails.ccnu.edu.cn, cgao@hust.edu.cn, liuyy@cug.edu.cn

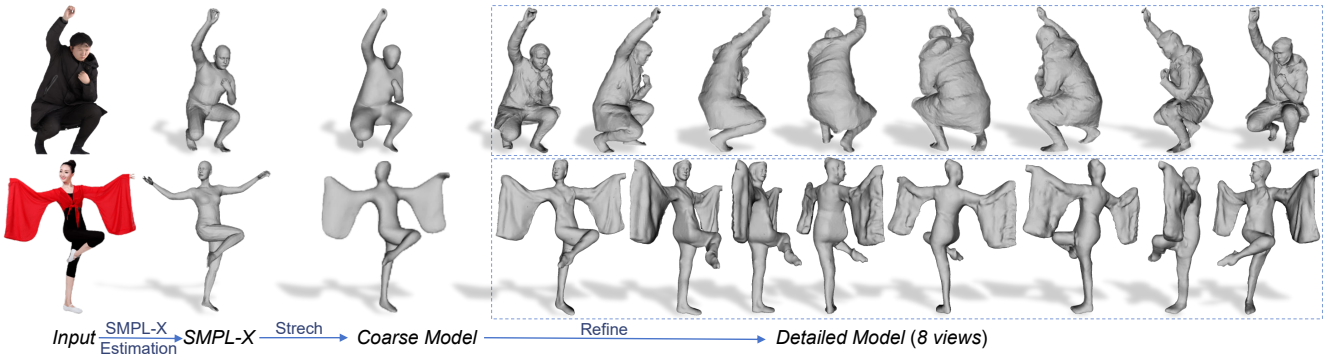

Figure 1. Using two stages of vertex shift, VS first stretches the estimated SMPL-X into a coarse human model and then refines it into a detailed clothed human model. Results on both benchmarks (1st row) and in-the-wild images (2nd row) show that VS can reconstruct *high-fidelity* and *artifact-less* clothed 3D humans from single images, even under scenarios of challenging poses and loose clothes. To comprehensively inspect results from all perspectives, we rotate the results through a full circle and show them from 8 views.

Abstract

Various applications require high-fidelity and artifact-free 3D human reconstructions. However, current implicit function-based methods inevitably produce artifacts while existing deformation methods are difficult to reconstruct high-fidelity humans wearing loose clothing. In this paper, we propose a two-stage deformation method named Vertex Shift (VS) for reconstructing clothed 3D humans from single images. Specifically, VS first stretches the estimated SMPL-X mesh into a coarse 3D human model using shift fields inferred from normal maps, then refines the coarse 3D human model into a detailed 3D human model via a graph convolutional network embedded with implicit-function-learned features. This "stretch-refine" strategy addresses large deformations required for reconstructing loose clothing and delicate deformations for recovering intricate and detailed surfaces, achieving high-fidelity reconstructions that faithfully convey the pose, clothing, and surface details from the input images. The graph convolutional network's ability to exploit neighborhood vertices coupled with the advantages inherited from the deformation methods ensure VS rarely produces artifacts like distortions and non-human shapes and never produces artifacts like holes, broken parts, and dismembered limbs. As a result, VS can reconstruct high-fidelity and artifact-less clothed 3D humans from single images, even under scenarios of challenging poses and loose clothing. Experimental results on three benchmarks and two in-the-wild datasets demonstrate that VS significantly outperforms current state-of-the-art methods. The code and models of VS are available for research purposes at https://github.com/starVisionTeam/VS.

1. Introduction

Human avatars will be an essential medium for connecting the physical and digital worlds. Recent advancements in 3D human reconstruction [7, 40] have made it convenient to create human avatars even from single phone-taken images, but the reconstructions are not yet perfect. An ideal reconstruction should be *high-fidelity* and *artifact-free*. More specifically, it should faithfully convey the pose, clothing, and surface details of the human in the image while being free from artifacts like holes, broken or missing parts, disembodied limbs, distortions, and non-human shapes.

Recently, encouraging reconstructions of humans with diverse clothing and poses have been yielded by implicit-function-based (IF-based) methods [14, 16, 36, 37]. They utilize deep neural networks to estimate occupancy fields [9, 31] or signed distance fields [32], from which detailed 3D human meshes can be reconstructed via MarchingCubes [29]. However, these IF-based methods often suffer from producing severe artifacts, as they do not exploit any knowledge about the human body structure. To address this issue, follow-up methods [3, 26, 48, 52, 54] utilize parametric body models [28, 34] to regularize the free-form implicit functions. While this regularization alleviates artifacts under the scenario of complex poses, it restricts the flexibility to loose clothing. The most recent work, ECON [49], has improved the quality of reconstructions to a new level, but it still produces reconstructions with wrong thickness, inaccurate poses, and missing parts in many cases. This indicates that more efforts are still needed to enhance the fidelity of the reconstructions and to eliminate artifacts. Besides IF-based methods, there is another line of works that use deformation methods, which naturally do not produce artifacts like holes and broken parts, for 3D object reconstruction [12, 13, 45]. Despite several works [22, 39, 47, 55] have applied deformation methods to clothed 3D human reconstruction, none of them has successfully reconstructed humans with loose clothing from single images. On the one hand, large deformations are required to fit loose clothing; on the other hand, recovering surface details needs to shift the mesh's vertices delicately.

In this paper, we propose *VS*, which stands for vertex shift, to reconstruct high-fidelity and artifact-less clothed 3D humans from single images. VS is a two-stage deformation method that uses a "stretch-refine" strategy to reconcile the contradiction between large deformations required for reconstructing loose clothing and delicate deformations for recovering details. Specifically, VS first stretches the estimated SMPL-X mesh [34] into a coarse 3D human model using displacement fields inferred from normal maps, then refines the coarse 3D human model into a detailed 3D human model via a graph convolutional network (GCN) embedded with implicit-function-learned features. In the stretching stage, VS shifts all vertices on the SMPL-X mesh to property positions where the coarse model can well align with the human in the input image. This enables VS to deal with loose clothing and correct pose/shape errors in the estimated SMPL-X. In the refining stage, the GCN driven by implicit-function-learned features can delicately deform the coarse model to recover the intricate and detailed surfaces of clothed humans. Moreover, each vertex on the graph model is constrained by its neighborhoods, which takes human body priors into account to suppress artifacts like distortions and non-human shapes. During the whole deforming process, VS only shifts vertices without altering edges, guaranteeing that it consistently produces holistic 3D human meshes without artifacts like holes, broken parts, and dismembered limbs. As a result, VS can reconstruct high-fidelity and artifact-less clothed 3D humans on both benchmark datasets and in-the-wild images even under scenarios of challenging poses and loose clothing (See Fig.1).

Our contributions are summarized as follows:

(1) We propose a two-stage deformation method that uses a "stretch-refine" strategy for clothed 3D human reconstruction, contributing to reconciling the contradiction between large deformations for reconstructing loose clothing and delicate formations for recovering surface details.

(2) We introduce displacement fields inferred from normal maps for stretching the coarse model to align well with the input image, allowing our deformation method to handle loose clothing and correct inaccurate pose estimates.

(3) We combine implicit-function-learned features with a graph convolutional network, making VS not only recover surface details but also suppress artifacts.

2. Related Work

Body-agnostic methods. Given the remarkable expressive capability of deep implicit functions [8–10, 31, 32] for arbitrary 3D object surfaces, mainstream body-agnostic methods [1, 14, 23, 36, 37] employ them to represent clothed 3D humans. PIFu [36] proposes pixel-aligned image features for training the implicit function. PIFuHD [37] further enhances the geometric details by introducing a multi-level architecture that leverages both low- and high-resolution images. While PIFu and PIFuHD can recover intricate local shapes of 3D humans, they often suffer from artifacts such as disembodied limbs and non-human shapes, especially in the case of complex or novel poses. To reduce artifacts, Geo-PIFu [14] predicts a voxel-based coarse geometry model to regularize the deep implicit function. However, for humans with novel poses that are not included in the training data, Geo-PIFu often predicts poor coarse geometry models. Besides geometry, Stereo-PIFu [16] and Self-portrait [23] utilize depth information to regularize implicit functions, while PHORHUM [1] considers shading and lighting information to infer better geometric details. Since body-agnostic methods do not exploit the human body structure, which is a strong prior for 3D human reconstruction, such methods often struggle with artifacts.

Body-aware methods. Body-aware methods [3, 5, 11, 15, 18, 19, 24–27, 42, 44, 48, 49, 52–54] utilize 3D parametric body models such as SMPL [28] and SMPL-X [34] or 2D joint points [6] as additional priors for clothed 3D human reconstruction. DeepHuman [53] fuses a semantic volume generated from the SMPL model with image features. However, due to an over-reliance on SMPL, DeepHuman is only able to reconstruct humans in tight clothing. PAMIR [54], HEI-Human [26], IP-Net [3], and DeepMult-

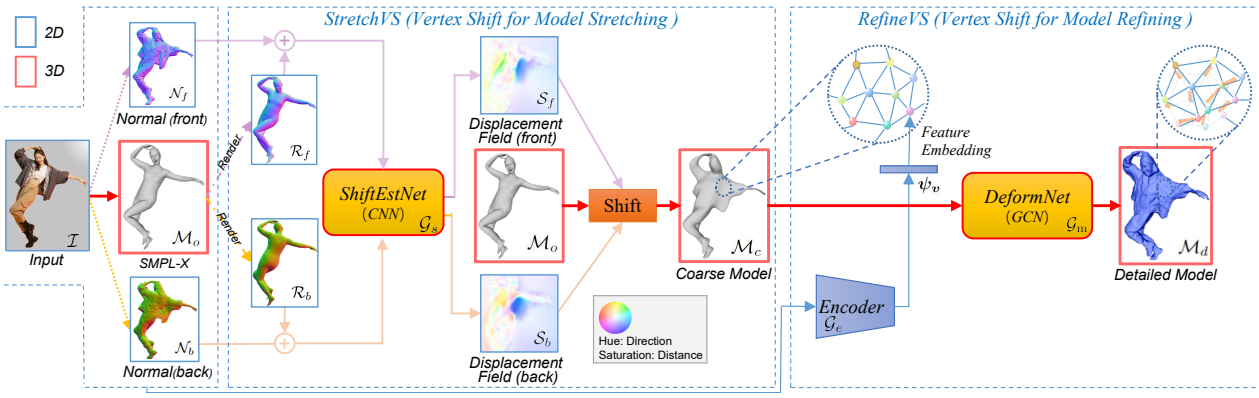

Figure 2. Overview of VS. VS employs a "stretch-refine" strategy to stepwise deform the SMPL-X ($\mathcal{M}_o$) into a coarse human model ($\mathcal{M}_c$) and a detailed human model ($\mathcal{M}_d$) using StretchVS and RefineVS modules, respectively. (1) First, two displacement fields ($\mathcal{S} = \{\mathcal{S}_f, \mathcal{S}_b\}$) are inferred by warping the body normals ($\mathcal{R} = \{\mathcal{R}_f, \mathcal{R}_b\}$) into clothing normal maps ($\mathcal{N} = \{\mathcal{N}_f, \mathcal{N}_b\}$) via a CNN (ShiftEstNet, $\mathcal{G}_s$). Then, StretchVS shifts vertices of the SMPL-X to from the coarse model using the displacement fields. (2) Taking the coarse model as input, RefineVS employs a GCN (DeformNet, $\mathcal{G}_{\mathrm{m}}$) embedded with implicit-function-learned features (Ψ) to infer vertex locations of the detailed model.

iCap [52] use voxelized SMPL to regularize free-form implicit functions. Nevertheless, this regularization restricts these methods' ability to reconstruct humans in loose clothing. Moreover, these methods are sensitive to the accuracy of SMPL estimates, and inaccurate SMPLs may introduce severe artifacts. ARCH [19], ARCH++ [15], FITE [25], and CAR [24] first learn the 3D shape of a human in the canonical space and then transform the learned shape into the posed space using SMPL. ICON [48] employs the SMPL model to guide both the detailed normals estimator and the visibility-aware implicit surface regressor. Recently, ECON [49] first utilizes BINI [4] to recover 2.5D front and back surfaces based on estimated normal maps, and then employs SMPL-X as a "canvas" to stitch together the two surfaces and "inpaint" the missing geometry. Although encouraging improvements have been achieved by these body-aware methods, they still inevitably produce artifacts and cannot yield high-fidelity reconstructions in many cases.

Deformation methods for 3D object reconstruction. Our work is also inspired by the deformation-based 3D object reconstruction methods [13, 21, 22, 39, 45]. Pix2Mesh [45] employs a graph-based neural network to progressively deform ellipsoid meshes into 3D object models. Point2Mesh [13] iteratively shrinks and wraps a convex hull of the input point cloud to a watertight mesh. GCMR [21] uses a Graph-CNN to directly regress the 3D location of the down-sampled SMPL vertices. MGT-Net [22] utilizes a vertex-based deformation representation to model human shapes and employs cascaded multi-scale graph transformation networks to generate topology-consistent 3D human models. Although holistic 3D models can be yielded, these methods are only suitable for reconstructing humans wearing tight clothing.

3. Method

An overview of VS is depicted in Fig. 2. Given an input image, VS first estimates an SMPL-X mesh and reorganizes it as a watertight 2-manifold graph (Sec. 3.1), then stretches the SMPL-X mesh into a coarse human mesh using displacement fields inferred from visual cues (Sec. 3.2), and finally refines the coarse human mesh by a GCN to form the final detailed human mesh (Sec. 3.3).

3.1. Model Representation

In VS, 3D meshes including the SMPL-X ($\mathcal{M}_o$), the coarse and detailed clothed human models ($\mathcal{M}_c$ and $\mathcal{M}_d$) are represented by graphs:

$$\mathcal{M}_* = \langle \mathcal{V}_*, \mathcal{E} \rangle \tag{1}$$

where $* = \{o, c, d\}$, $\mathcal{V}_* = \{\boldsymbol{v}_*^1, \boldsymbol{v}_*^2, \cdots\}$ are 3D vertices, and $\mathcal{E} = \{e^1, e^2, \cdots\}$ are edges. Particularly, we reorganize the vertices and edges to form watertight 2-manifold triangular meshes using [17]. The manifold and watertight properties ensure that our meshes have consistent surfaces without any holes, broken parts, or disembodied limbs, which are often observed in meshes reconstructed by IF-based methods. As mentioned above, VS stepwise deforms the SMPL-X model into coarse and detailed human models. In order to preserve the manifold and watertight properties of meshes, VS only shifts vertices without altering edges during the deforming process.

The original SMPL-X[34] only has 10,475 vertices and 20,908 faces, and meshes deformed from it are challenging to accurately represent clothed 3D humans, especially when loose clothing is involved. To address this issue, we use the method proposed in [17] to increase the vertices and faces of the SMPL-X mesh.

3.2. Vertex Shift for Coarse Model Stretching

The StretchVS module stretches the SMPL-X mesh ($\mathcal{M}_o$) into the coarse human mesh ($\mathcal{M}_c$) using displacement fields ($\mathcal{S} = \{\mathcal{S}_f, \mathcal{S}_b\}$ inferred from visual cues:

$$\text{Shift}(\mathcal{M}_o, \mathcal{S}) \rightarrow \mathcal{M}_c \Leftrightarrow \mathcal{V}_\text{o} + \mathcal{V}_\text{s} \rightarrow \mathcal{V}_\text{c} \tag{2}$$

where $\mathcal{V}_\text{o}$ and $\mathcal{V}_\text{c}$ are respectively the vertices of $\mathcal{M}_o$ and $\mathcal{M}_c$, and $\mathcal{V}_\text{s} = \{\boldsymbol{v}_s^1, \boldsymbol{v}_s^2, \cdots\}$ is a set of shift vectors derived from $\mathcal{S}$. For each vertex $\boldsymbol{v}_s^i$ in $\mathcal{V}_s$:

$$\boldsymbol{v}_s^i = \begin{cases} \left\langle \mathcal{S}_f(\pi(\boldsymbol{v}_o^i)), 0 \right\rangle, & \text{if } \boldsymbol{v}_o^i \text{ is visible} \\ \left\langle \mathcal{S}_b(\pi(\boldsymbol{v}_o^i)), 0 \right\rangle, & \text{else} \end{cases} \tag{3}$$

where and $\pi(\cdot)$ represents the orthogonal transformation that projects a 3D vertex onto the 2D displacement field plane. As described in Equ. 3, StretchVS shifts vertices on X-Y planes while maintaining the depth of SMPL-X, since it is challenging to estimate high-precision depth from 2D images using concise algorithms. Vertex shifts in the Z-direction will be handled by the RefineVS module later.

The key to StretchVS is inferring the displacement fields from visual cues. Pioneering works like PIFuHD [37], ICON [48] and ECON [49] have already demonstrated the effectiveness of normals in clothed 3D human reconstruction. Encouraged by these works, we choose normals as the visual cues for inferring the displacement fields.

Normals generation. We employ PyMAF-X [51] to estimate the SMPL-X, and then render two body normal maps ($\mathcal{R} = \{\mathcal{R}_f, \mathcal{R}_b\}$) from the SMPL-X respectively from the front and back views. To estimate clothing normal maps ($\mathcal{N} = \{\mathcal{N}_f, \mathcal{N}_b\}$) from the input image, the approach presented in ECON [49] is used.

Displacement field inferring. In our method, the displacement fields are inferred by establishing the pixel-wise correspondences [41] between the body normal maps ($\mathcal{R}$) and the clothing normal maps ($\mathcal{N}$). Once $\mathcal{S}$ is inferred, $\mathcal{R}$ can be warped into $\mathcal{N}$:

$$\hat{\mathcal{N}_*} = \mathcal{W}(\mathcal{R}_*, \mathcal{S}_*) \tag{4}$$

where $* = \{f, b\}$, and $\mathcal{W}(\cdot)$ is the dense image warp function [2]. For simplicity, we omit the subscripts 'f' and 'b' for the front and back views in the subsequent formulas.

We design a 5-layer lightweight CNN, named ShiftEstNet ($\mathcal{G}_s$), to infer the displacement fields ($\mathcal{S}$):

$$\mathcal{G}_s(\mathcal{R}, \mathcal{N}) \rightarrow \mathcal{S} \tag{5}$$

Since the body normals and the clothing normals lay in different domains, there is no direct photometric evidence for establishing the dense correspondences between them. Inspired by [56] and [41], we propose a "cycle wraping" strategy. To this end, ShiftEstNet is designed to have the capacity of inverse warping:

$$\mathcal{G}_s(\mathcal{N}, \mathcal{R}) \rightarrow \tilde{\mathcal{S}} \tag{6}$$

where $\tilde{\mathcal{S}}$ is the inferred displacement fields for inverse warping:

$$\hat{\mathcal{R}} = \mathcal{W}(\mathcal{N}, \tilde{\mathcal{S}}) \tag{7}$$

In a cycle, we can warp body normal maps into the clothing normal maps and then warp them back to body normal maps, and vice versa:

$$\begin{cases} \hat{\mathcal{R}}_c = \mathcal{W}\left(\mathcal{W}(\mathcal{R}, \mathcal{S}), \tilde{\mathcal{S}}\right) \\ \hat{\mathcal{N}}_c = \mathcal{W}\left(\mathcal{W}(\mathcal{N}, \tilde{\mathcal{S}}), \mathcal{S}\right) \end{cases} \tag{8}$$

In this way, ShiftEstNet can be optimized in an self-supervised manner. The objective function consists of three items:

$$\min_{\mathcal{S}, \tilde{\mathcal{S}}} \lambda_\text{w}\mathcal{L}_\text{w} + \lambda_\text{c}\mathcal{L}_\text{c} + \lambda_\text{s}\mathcal{L}_\text{s} \tag{9}$$

where $\mathcal{L}_\text{w}$ is a warping loss:

$$\mathcal{L}_\text{w} = \|\mathcal{N} - \hat{\mathcal{N}}\|_2^2 + \|\mathcal{R} - \hat{\mathcal{R}}\|_2^2 \tag{10}$$

$\mathcal{L}_\text{c}$ is a cycle consistent loss:

$$\mathcal{L}_\text{c} = \|\mathcal{N} - \hat{\mathcal{N}}_c\|_2^2 + \|\mathcal{R} - \hat{\mathcal{R}}_c\|_2^2 \tag{11}$$

and $\mathcal{L}_\text{s}$ is a smoothness loss that encourages the estimated displacement fields to be locally smooth [38]:

$$\mathcal{L}_\text{s} = \|\mathcal{N}_g - \hat{\mathcal{N}}_g\|_2^2 \tag{12}$$

$$\hat{\mathcal{N}}_g = \mathcal{W}\left(\mathcal{W}(\mathcal{N}_g, \tilde{\mathcal{S}}), \mathcal{S}\right) \tag{13}$$

where $\mathcal{N}_g$ is an uniform-distributed random normal map.

3.3. Vertex Shift for Detailed Model Refining

Driven by embedded features (Ψ), the RefineVS module deforms the coarse human mesh ($\mathcal{M}_c$) into the detailed human mesh ($\mathcal{M}_d$) via a GCN named DeformNet ($\mathcal{G}_\text{m}$):

$$\mathcal{G}_\text{m}\left(\mathcal{M}_c; \Psi\right) \rightarrow \mathcal{M}_d \tag{14}$$

Network structure. DeformNet is designed using the mesh convolution operation proposed in MeshCNN [12]. Recall that $\mathcal{M}_c$ is a watertight 2-manifold mesh with triangular faces, which guarantees that each edge in the mesh is exactly adjacent to four other edges. Leveraging this invariant, we apply convolution operations to edges rather than vertices. The convolution for an edge e_0 and its four adjacent edges $\{e_1, e_2, e_3, e_4\}$ is defined as:

$$\sum_{k=0}^{4} \boldsymbol{\mu}_k \cdot \boldsymbol{\psi}_{e_k} \tag{15}$$

where $[\boldsymbol{\mu}_0, \cdots, \boldsymbol{\mu}_4]$ are convolutional kernels, and $\boldsymbol{\psi}_{e_k}$ is the feature embedded on edge e_k.

$$\boldsymbol{\psi}_{e_k} = \boldsymbol{\psi}_{v_i} \oplus \boldsymbol{\psi}_{v_j} \tag{16}$$

where v_i and v_j are the two vertices of edge e_k, and $\oplus$ denotes the concatenate operator.

DeformNet is designed as a fully convolutional network, so that intermediate layers and the output layer (i.e., the detailed mesh $\mathcal{M}_d$) after mesh convolution operations can maintain the same structure as the input mesh $\mathcal{M}_c$. In our implementation, DeformNet consists of 11 mesh convolutional layers, each followed by a LeakyReLU activation. Particularly, the number of channels of the convolution kernel before the output layer is set to 6, allowing us to use the 6-dimensional convolutional result to represent the positions of the two vertices of an edge. Note that a vertex is shared by an uncertain number of edges, which means that DeformNet predicts multiple new positions for each vertex. We take the mean of the multiple positions as the final positions of the vertex.

Features. From the seminal PIFu [36] to the current ECON [49], it has been demonstrated that "pixel-aligned" features trained using deep implicit functions achieve encouraging results in clothed 3D human reconstruction. In our method, we employ an encoder (denoted as $\mathcal{G}_e$) with a similar structure to PIFuHD [37] to extract visual features from the input image ($\mathcal{I}$) and the estimated normal maps ($\{\mathcal{N}_f, \mathcal{N}_b\}$):

$$\mathcal{G}_e\left(\mathcal{I}, \mathcal{N}_f, \mathcal{N}_b\right) \rightarrow \Psi \tag{17}$$

Although the encoder can be retrained, we have observed that using an off-the-shelf encoder is a more convenient and feasible solution. Thereby, we use the pre-trained encoder from PIFuHD [37]. Experiments show that DeformNet is not sensitive to feature encoders, and replacing the encoder with that of other methods like PIFu [36] and PaMIR [54] does not significantly decrease the quality of reconstructions (see "ablation study" in Sec. 4.4). In addition to visual features extracted by the encoder, we also embed the coordinate ($\boldsymbol{v}$) and normal ($\mathcal{N}_{\boldsymbol{v}}$) of each vertex onto the vertex:

$$\boldsymbol{\psi}_{\boldsymbol{v}} = \langle \Psi(\pi(\boldsymbol{v})), \mathcal{N}_{\boldsymbol{v}}, \boldsymbol{v} \rangle \tag{18}$$

where $\pi(\boldsymbol{v})$ denotes the orthogonal projection that maps a 3D vertex $\boldsymbol{v}$ onto the 2D feature plane.

Losses. The coarse model $\mathcal{M}_c$ input into DeformNet already has an approximate geometric topology, but it is not accurate enough and lacks local details. In addition, we expect not to introduce "bulge" artifacts when deforming $\mathcal{M}_c$ into $\mathcal{M}_d$. Thus, DeformNet faces three tasks: (1) improving geometry accuracy, (2) recovering local details, and (3) preventing "bulge" artifacts. To these ends, the loss for training DeformNet consists of a geometric error term ($\mathcal{L}_\mathrm{d}$), a local detail error term ($\mathcal{L}_\mathrm{n}$), and a regularization term ($\mathcal{L}_\mathrm{e}$):

$$\mathcal{L} = \lambda_\mathrm{d}\mathcal{L}_\mathrm{d} + \lambda_\mathrm{n}\mathcal{L}_\mathrm{n} + \lambda_\mathrm{e}\mathcal{L}_\mathrm{e} \tag{19}$$

Here, $\mathcal{L}_\mathrm{d}$ is described by Chamfer distance:

$$\mathcal{L}_\mathrm{d} = \sum_{\hat{\boldsymbol{p}} \in \hat{\mathcal{P}}} \min_{\boldsymbol{p} \in \mathcal{P}} \|\hat{\boldsymbol{p}} - \boldsymbol{p}\|_2 + \sum_{\boldsymbol{q} \in \mathcal{P}} \min_{\hat{\boldsymbol{q}} \in \hat{\mathcal{P}}} \|\hat{\boldsymbol{q}} - \boldsymbol{q}\|_2 \tag{20}$$

where $\hat{\mathcal{P}}$ and $\mathcal{P}$ are 3D point sets respectively sampled from the predicted ground-truth meshes. Note that, we sample uniformly distributed 3D points using the "triangle point picking" approach [46] when computing $\mathcal{L}_\mathrm{d}$, since the original vertices of the predicted and ground-truth meshes are usually not distributed uniformly on their surfaces. $\mathcal{L}_\mathrm{n}$ is described using the cosin distance of normals:

$$\mathcal{L}_\mathrm{n} = \sum_{\boldsymbol{p}, \hat{\boldsymbol{p}}} -\frac{\boldsymbol{n}_{\boldsymbol{p}} \cdot \boldsymbol{n}_{\hat{\boldsymbol{p}}}}{\|\boldsymbol{n}_{\boldsymbol{p}}\|_2 \|\boldsymbol{n}_{\hat{\boldsymbol{p}}}\|_2} + \sum_{\boldsymbol{q}, \hat{\boldsymbol{q}}} -\frac{\boldsymbol{n}_{\boldsymbol{q}} \cdot \boldsymbol{n}_{\hat{\boldsymbol{q}}}}{\|\boldsymbol{n}_{\boldsymbol{q}}\|_2 \|\boldsymbol{n}_{\hat{\boldsymbol{q}}}\|_2} \tag{21}$$

where $(\boldsymbol{p}, \hat{\boldsymbol{p}})$ and $(\boldsymbol{q}, \hat{\boldsymbol{q}})$ are bi-directional matched closest point pairs, and $\boldsymbol{n}$ represents the normal vector. The regularization term $\mathcal{L}_\mathrm{e}$ penalizes long edges:

$$\mathcal{L}_\mathrm{e} = \sum_{e \in \mathcal{E}} \|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2 \tag{22}$$

where $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are the two vertices of edge e. $\mathcal{L}_\mathrm{e}$ not only prevents "bulges" artifacts, but also helps convergence.

4. Experimental Results

VS is compared with five state-of-the-art (SOTA) single-image clothed 3D human reconstruction methods with publicly released code: PIFu [36], PIFuHD [37], PaMIR [54], ICON [48], and ECON [49]. We opt for the ECON_{EX} version of ECON [49], as it is reported to have superior and stable performance over the ECON_{IF} version [49]. All methods employ the post-processing technique used in ICON. For a fair comparison, we use the re-implemented PIFu and PaMIR provided by Xiu et al. [48], which have uniformized network settings and inputs with ICON. As such, all methods involved in our experiments fairly take an RGB image and the estimated front/back normal maps as input.

4.1. Datasets

Training dataset. As with ECON [49], VS is trained on the THuman2.0 dataset [50], which contains 525 3D scans and SMPL-X models of 525 subjects with lab-acted poses and daily clothes. In our experiments, we split THuman2.0 into a training set and a testing set at a ratio of 20:1 by subjects. To augment the training data, we render 18 RGB images for each 3D scan by rotating a virtual camera w.r.t the frontal view at an interval of 20^o. In the end, 9K pairs (RGB image, 3D scan) are obtained for training.

Testing datasets. Our VS is extensively evaluated on five datasets: the testing set of THuman 2.0 [50], CAPE [30], RenderPeople [35], SHHQ [20], and VcgPeople. Humans in these testing datasets exhibit a variety of poses and clothes, many of which are novel in the training data. For THuman2.0, CAPE, and RenderPeople, which have ground-truth 3D scans, we render 6 testing images from each scan at random viewpoints. Since SHHQ and VcgPeople just contain in-the-wild images, we only conduct qualitative evaluations on them.

Methods	Publications	THuman 2.0				CAPE				RenderPeople			
		$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$
PIFu [36]	ICCV'19	1.760	1.904	0.0500	0.2408	2.967	2.738	0.0449	0.2409	2.781	2.857	0.0590	0.2773
†PIFuHD [37]	CVPR'20	3.088	3.113	0.0891	0.3663	4.714	3.823	0.0555	0.2796	3.311	3.3118	0.0846	0.3541
PaMIR [54]	TPAMI'22	1.064	1.185	0.0438	0.1927	1.772	1.404	0.0337	0.1676	1.580	1.659	0.0486	0.2088
ICON [48]	CVPR'22	0.947	0.925	0.0422	0.1761	1.133	1.096	0.0311	0.1431	1.265	1.251	0.0431	0.1871
ECON [49]	CVPR'23	0.906	0.845	0.0379	0.1891	0.937	0.921	0.0335	0.1644	1.285	1.079	**0.0417**	0.1973
VS	Ours	**0.628**	**0.555**	**0.0373**	**0.1555**	**0.621**	**0.615**	**0.0262**	**0.1138**	**0.976**	**0.788**	0.0419	**0.1618**

Table 1. Quantitative comparison against other SOTAs on three benchmarks. Ground-truth SMPLs are used for methods that need SMPL-(x) guidance. † Official model trained on Renderpeople.

4.2. Quantitative Evaluation

Metrics. Chamfer distance (ε_{cd}, unit: cm), point-to-surface distance (ε_{p2s}, unit: cm), cosine distance ($\varepsilon_{cos} \in [0, 1]$), and L2 normal error ($\varepsilon_{l2} \in [0, 1]$) are employed as quantitative metrics. The first two metrics mainly assess large geometric errors, while the latter two tend to measure local detail differences between reconstructions and ground-truth models. Following ECON [49], when reporting cosine distance and L2 normal error, four normal maps are rendered respectively for the reconstructed and ground-truth models at the angles of $\{0^o, 90^o, 180^o, 270^o\}$.

Quantitative comparison. As illustrated in Table 1, VS demonstrates superior performance on all three benchmark datasets. Especially, our VS exceeds all other methods by a large margin on the two metrics that represent large geometric errors (i.e., ε_{cd} and ε_{p2s}). In terms of local surface details (i.e., ε_{cos} and ε_{L2}), VS performs on par with ECON and much better than PIFu, PIFuHD, PaMIR, and ICON.

4.3. Qualitative Evaluation

Visual results. Fig. 4 shows the reconstructions by applying our VS on in-the-wild images. It can be seen that VS recovers high-fidelity clothed 3D humans even in challenging poses and complex loose clothes. Note that the poses, clothes, and accessories in these images are out of the distribution of the training data. These results reveal that VS not only has high reconstruction accuracy on challenging images but also has strong generalization capabilities.

Visual comparison. Given that ECON [49] is the current SOTA method for single-image clothed 3D human reconstruction, we first compare our VS with it solely. Fig. 3 shows the reconstructions by ECON and our VS. It can be seen that VS recovers high-fidelity 3D humans while ECON struggles with missing parts, incomplete loose clothing, wrong thickness, and incorrect poses. Fig. 5 demonstrates visual comparisons with all five SOTA methods on in-the-wild images. ECON and VS demonstrate a substantial advantage over PIFu [36], PIFuHD [37], PaMIR [54], and ICON [48].

Perceptual study. We also conduct a perceptual study on the in-the-wild SHHQ and VcgPeople datasets. To this end, we designed a perceptual evaluation tool (see Fig. 11 in SupMat) where reconstructions produced by different methods are blindly placed on one page. We recruited 25 volunteers to participate in the perceptual study, and randomly sampled 50 images from each of the two datasets for each participant. Participants were requested to select the ***best*** reconstruction for each input image after rotating the reconstructions and carefully inspecting them from all viewpoints. The average percentages of preferences from these participants are illustrated in Table 2. The majority of participants perceived that our VS achieves the best reconstructions on both datasets. It's worth noting that ECON and VS exhibit greater superiorities over other methods on the more challenging VcgPeople dataset.

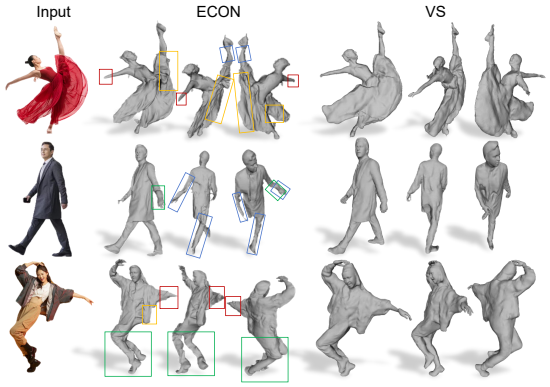

Figure 3. Comparison of VS with ECON on in-the-wild images. VS recovers high-fidelity 3D humans while ECON struggles with missing parts, incomplete loose clothing, wrong thickness, and incorrect poses. More comparisons are shown in Video2 in SupMat.

Datasets	PIFu	PIFuHD	PaMIR	ICON	ECON	VS
SHHQ	3.6	4.3	11.1	3.9	19.0	**58.1**
VcgPeople	0.2	0.1	5.6	5.0	24.7	**64.4**

Table 2. Average percentages (%) of participants' preferences. VS is significantly preferred over other methods.

4.4. Ablation Study

Effectiveness of the two vertex shift states. To assess the effectiveness of the StretchVS and RefineVS modules in VS, we implemented two variants of VS, each exclusively incorporating either the StretchVS module or the RefineVS module. The quantitative results achieved by the complete

Figure 4. Results produced by VS on in-the-wild images. Each result is shown from three views. VS recovers high-fidelity and artifact-less clothed 3D humans even in challenging poses (P) and loose clothes (C). More results are shown in Fig.12 and Video1 in SupMat.

Versions	THuman 2.0				CAPE				RenderPeople			
	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$
VS	0.628	0.555	0.0373	0.1555	0.621	0.615	0.0262	0.1138	0.976	0.788	0.0419	0.1618
StretchVS	0.979	0.915	0.0645	0.2469	0.935	0.907	0.0404	0.1743	1.209	1.040	0.0627	0.2443
RefineVS	0.669	0.598	0.0391	0.1679	0.721	0.719	0.0271	0.1293	1.263	0.997	0.0529	0.2138

Table 3. Quantitative results achieved by three versions of VS.

VS and the two variants are presented in Table 3. It is evident from these results that: (1) StretchVS exhibits a significant performance decrease on all three datasets, especially in terms of local details. (2) RefineVS performs consistently on THuman2.0 and CAPE without significant degradation, while it exhibits a noticeable decline on RenderPeople. This is because RenderPeople contains many out-of-distribution loose clothes, and it is difficult for the RefineVS module to deform an SMPL-X mesh to fit loose clothes without the assistance of the StretchVS module. The visual results shown in Fig. 6 are mutually corroborated with the quantitative results illustrated in Table 3.

Robustness to SMPL-X estimates. Most body-aware methods are known to be sensitive to SMPL-(X) estimates, while SMPL-(X) estimation remains an unresolved issue. In this experiment, we use both accurate and inaccurate SMPL-X estimates to initiate VS. As shown in Fig. 7, VS can achieve high-quality reconstructions even with inaccurate SMPL-X estimates. Since StretchVS treats SMPL-X just as a "seed" and stretches it to align with visual cues, it has the "error-correcting" ability to shift vertices on inaccurate SMPL-X estimates to proper positions.

Flexibility to feature encoders. To validate that VS is insensitive to IF-based feature encoders, we evaluated VS

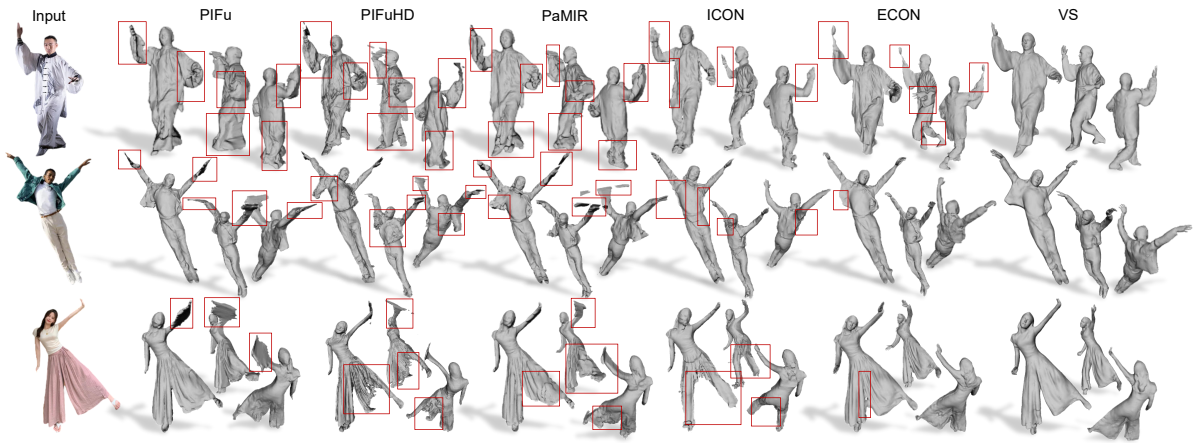

Figure 5. Visual comparison with other SOTA methods on in-the-wild images. More results are shown in Fig.16∼19 in SupMat.

Encoders	THuman 2.0				CAPE				RenderPeople			
	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$
PIFu [36]	0.766	0.688	0.0411	0.1790	0.735	0.736	0.0272	0.1291	1.104	0.884	0.0465	0.1939
PaMIR [54]	0.838	0.733	0.0402	0.1768	0.861	0.799	0.0266	0.1310	1.182	0.933	0.0428	0.1879
PIFuHD [37]	0.628	0.555	0.0373	0.1555	0.621	0.615	0.0262	0.1138	0.976	0.788	0.0419	0.1618

Table 4. Quantitative evaluation of VS with different feature encoders.

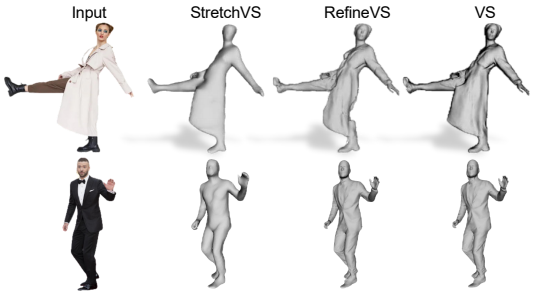

Figure 6. Results produced by VS variants w/wo StretchVS and RefineVS modules. StretchVS handles large deformations like loose clothing, while RefineVS recovers surface details. For humans who don't wear loose clothing, RefineVS can work alone.

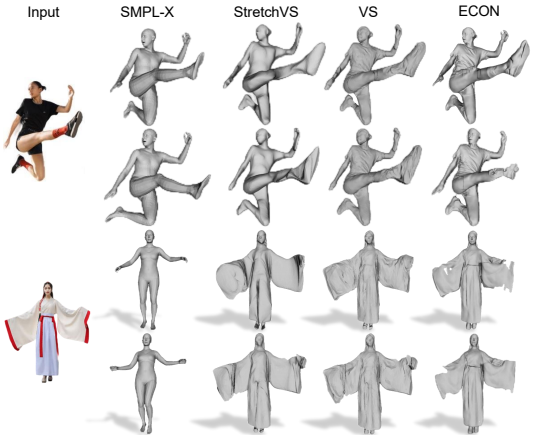

Figure 7. Robustness to SMPL-X estimates. We use both accurate (odd rows) and inaccurate (even rows) SMPL-X estimates to initiate VS. VS can achieve high-quality reconstructions even with inaccurate SMPL-X estimates.

with pre-trained encoders from PIFu [36], PaMIR [54], and PIFuHD [37]. As shown in Table 4, the employment of different feature encoders does not result in significant discrepancies in the quantitative metrics achieved by VS.

5. Conclusion

We propose VS to reconstruct high-fidelity and artifact-less clothed 3D humans from single images. VS first stretches the estimated SMPL-X model into a coarse 3D human model using displacement fields inferred from visual cues, and then refines the coarse 3D human model into a detailed 3D human model via a GCN embedded with implicit-function-learned features. This two-stage deformation strategy reconciles the contradiction between large deformations for reconstructing loose clothing and delicate vertex shifts for recovering intricate and detailed surfaces. Extensive experiments on five datasets demonstrate that VS can reconstruct high-fidelity and artifact-less clothed 3D humans and achieves SOTA performance. VS confirms that deformation methods can reconstruct high-quality clothed 3D humans with complex poses and loose clothing, and even have advantages over IF-based methods in eliminating artifacts. We hope that VS can provide a new perspective for 3D human reconstruction and bring us one step closer to creating "perfect" human avatars for real-world applications.

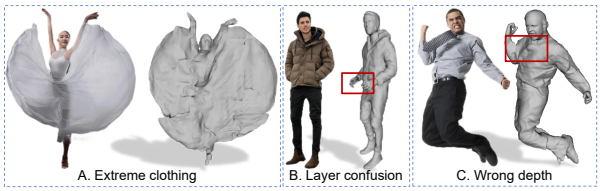

Figure 8. Failure cases. VS may fail to handle extreme clothing, layer confusion, and large deformation in the Z-axis.

Limitations & Future work. Although VS achieves impressive results, it still has some limitations, as shown in Fig. 8. (A) Extreme clothing. VS may face challenges in deforming vertices on the legs of SMPL-X into extreme dresses. (B) Layer confusion. Currently, VS cannot distinguish the layers of the human body and clothing. (C) Wrong depth. VS fails in a couple of images exhibiting large deformation in the Z-axis (say, belts drifting towards the camera). In future work, we plan to separate a clothed human into the body and clothing parts and reconstruct them separately. As the clothing styles can be identified from the images [43] and the corresponding preset models can be provided [33], we hypothesize that using VS to deform preset clothing and body models separately can reconstruct extreme clothing and mitigate layer confusion.

6. Acknowledgment

This work was supported by the National Natural Science Foundation of China (62077026), and the Fundamental Research Funds for the Central Universities (CCNU22QN012).

References

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1506–1515, 2022. 2

[2] The TensorFlow Authors. Dense image warp. www.tensorflow.org/addons/api_docs/python/tfa/image/dense_image_warp, 2023. 4

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Eur. Conf. Comput. Vis. (ECCV)*, page 311–329, 2020. 2

[4] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 552 – 567, 2022. 3

[5] Yukang Cao, Kai Han, and Kwan-Yee K. Wong. Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4647–4657, 2023. 2

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 7291–7299, 2017. 2

[7] Lu Chen, Sida Peng, and Xiaowei Zhou. Towards efficient and photorealistic 3d human reconstruction: A brief survey. *Visual Informatics*, 5(4):11–19, 2021. 1

[8] Weikai Chen, Cheng Lin, Weiyang Li, and Bo Yang. 3psdf: Three-pole signed distance function for learning surfaces with arbitrary topologies. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 18522–18531, 2022. 2

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5932–5941, 2019. 2

[10] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Int. Conf. Mach. Learn. (ICML)*, pages 3789–3799, 2020. 2

[11] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 12869–12879, 2023. 2

[12] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes-Or, Shachar Fleishman, and Daniel. MeshCNN: A network with an edge. *ACM Trans. Graph. (TOG)*, 38(4):90:1–90:12, 2019. 2, 4

[13] Rana Hanocka, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Point2mesh: A self-prior for deformable meshes. *ACM Trans. Graph. (TOG)*, 39(4):126:1–126:12, 2020. 2, 3, 1

[14] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 2304–2314, 2020. 2

[15] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-ready clothed human reconstruction revisited. In *Int. Conf. Comput. Vis. (ICCV)*, pages 11046–11056, 2021. 2, 3

[16] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. StereoPIFu: Depth aware clothed human digitization via stereo vision. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 119–128, 2021. 2

[17] Jingwei Huang, Hao Su, and Leonidas J. Guibas. Robust watertight manifold surface generation method for shapenet models. *ArXiv*, 2018. 3, 1

[18] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *Int. Conf. 3D Vis.(3DV)*, 2024. 2, 1

[19] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH:animatable reconstruction of clothed humans. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3093–3102, 2020. 2, 3

[20] Fu Jianglin, Li Shikai, Jiang Yuming, Lin Kwan-Yee, Qian Chen, Loy Chen Change, Wu Wayne, and Liu Ziwei. Stylegan-human: A data-centric odyssey of human generation. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 1–19, 2022. 5

[21] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4501–4510, 2019. 3

[22] Kun Li, Wen Hao, Feng Qiao, Yuxiang Zhang, Xiongzheng Li, Jing Huang, Cunkuan Yuan, Yu-Kun Lai, and Yebin Liu. Image-guided human reconstruction via multi-scale graph transformation networks. *IEEE Trans. Image Process. (TIP)*, 30:5239–5251, 2021. 2, 3

[23] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3d self-portraits in seconds. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1344–1353, 2020. 2

[24] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-fidelity clothed avatar reconstruction from a single image. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 8662–8672, 2023. 2, 3

[25] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *Eur. Conf. Comput. Vis. (ECCV)*, page 210–228, 2022. 3

[26] Leyuan Liu, Jianchi Sun, Yunqi Gao, and Jingying Chen. HEI-Human: A hybrid explicit and implicit method for single-view 3d clothed human reconstruction. In *Chinese Conf. Pattern Recog. Comput. Vis. (PRCV)*, pages 251–262, 2021. 2

[27] Leyuan Liu, Yunqi Gao, Jianchi Sun, and Jingying Chen. Single-image clothed 3d human reconstruction guided by a well-aligned parametric body model. *Multimedia Systems*, page 1579–1592, 2023. 2

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph. (TOG)*, 34(6):1–16, 2015. 2

[29] William Lorensen and Harvey Cline. Marching Cubes: A high resolution 3d surface construction algorithm. *ACM Trans. Graph. (TOG)*, 21(4):163–169, 1987. 2

[30] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3d people in generative clothing. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6469–6478, 2020. 5, 1

[31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Networks: Learning 3d reconstruction in function space. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4455–4465, 2019. 2

[32] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 165–174, 2019. 2

[33] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 7365–7375, 2020. 8

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 10975–10985, 2019. 2, 3

[35] RenderPeople. www.renderpeople.com, 2018. 5, 1

[36] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Int. Conf. Comput. Vis. (ICCV)*, pages 2304–2314, 2019. 2, 5, 6, 8, 1

[37] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 84–93, 2020. 2, 4, 5, 6, 8, 1

[38] Siyuan Shan, Wen Yan, Xiaoqing Guo, Eric I-Chao Chang, Yubo Fan, and Yan Xu. Unsupervised end-to-end learning for deformable medical image registrations. *ArXiv*, 2018. 4

[39] Qingyang Tan, Lin Gao, Yu-Kun Lai, Jie Yang, and Shihong Xia1. Mesh-based autoencoders for localized deformation component analysis. In *AAAI*, page 2452–2459, 2018. 2, 3

[40] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3D human mesh from monocular images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, pages 15406 – 15425, 2023. 1

[41] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *Int. Conf. Comput. Vis. (ICCV)*, pages 10346–10356, 2021. 4

[42] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 20–36, 2018. 2

[43] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling cross stage partial network. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 13029–13038, 2021. 8

[44] Junying Wang, Jae Shin Yoon, Tuanfeng Y. Wang, Krishna Kumar Singh, and Ulrich Neumann. Complete 3d human reconstruction from a single incomplete image. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 8748–8758, 2023. 2

[45] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3d mesh models from single rgb images. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 52–67, 2018. 2, 3

[46] Eric W. Weisstein. Triangle point picking. mathworld.wolfram.com/TrianglePointPicking.html, 2023. 5

[47] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *Int. Conf. 3D Vis.(3DV)*, pages 322–332, 2020. 2

[48] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. ICON: Implicit clothed humans obtained from normals. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 13296–13306, 2022. 2, 3, 4, 5, 6, 1

[49] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit clothed humans optimized via normal integration. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 512–523, 2023. 2, 3, 4, 5, 6, 1

[50] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4D: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5746–5756, 2021. 5, 1

[51] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: towards well-aligned full-body model regression from monocular images. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 45(10):12287 – 12303, 2023. 4

[52] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. DeepMultiCap: Performance capture of multiple characters using sparse multiview cameras. In *Int. Conf. Comput. Vis. (ICCV)*, pages 6219–6229, 2021. 2, 3

[53] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3d human reconstruction from a single image. In *Int. Conf. Comput. Vis. (ICCV)*, pages 7739–7749, 2019. 2

[54] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 44(6):3170 – 3184, 2022. 2, 5, 6, 8, 1

[55] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4491–4500, 2019. 2

[56] Jun-yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis. (ICCV)*, pages 2223–2232, 2017. 4