SeSDF: Self-evolved Signed Distance Field for Implicit 3D Clothed Human Reconstruction

Yukang Cao   Kai Han   Kwan-Yee K. Wong

The University of Hong Kong

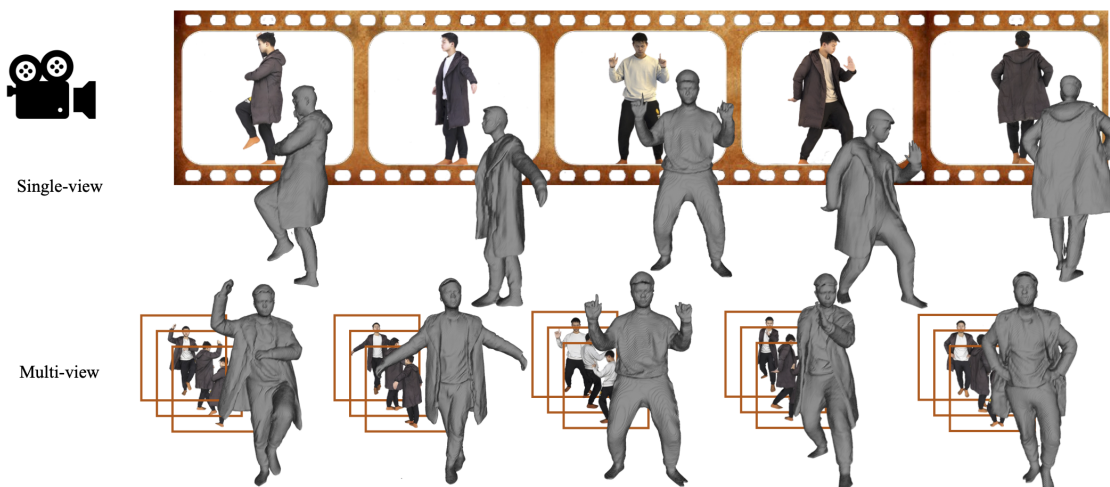{ykcao, kykwong}@cs.hku.hk   kaihanx@hku.hk

Figure 1. **Self-evolved Signed Distance Field (SeSDF):** We propose Self-evolved Signed Distance Field (SeSDF), which can flexibly reconstruct 3D clothed human models from a single-view image (top) or uncalibrated multi-view images (bottom). SeSDF can robustly recover fine geometry details from any type of poses, which allows us to generate clothed human avatars.

Abstract

We address the problem of clothed human reconstruction from a single image or uncalibrated multi-view images. Existing methods struggle with reconstructing detailed geometry of a clothed human and often require a calibrated setting for multi-view reconstruction. We propose a flexible framework which, by leveraging the parametric SMPL-X model, can take an arbitrary number of input images to reconstruct a clothed human model under an uncalibrated setting. At the core of our framework is our novel self-evolved signed distance field (SeSDF) module which allows the framework to learn to deform the signed distance field (SDF) derived from the fitted SMPL-X model, such that detailed geometry reflecting the actual clothed human can be encoded for better reconstruction. Besides, we propose a simple method for self-calibration of multi-view images via the fitted SMPL-X parameters. This lifts the requirement of tedious manual calibration and largely increases the flexibility of our method. Further, we introduce an effective occlusion-aware feature fusion strategy to account for the most useful fea-
tures to reconstruct the human model. We thoroughly evaluate our framework on public benchmarks, demonstrating significant superiority over the state-of-the-arts both qualitatively and quantitatively.

1. Introduction

Clothed human reconstruction is a hot topic with increasing demand in real-world applications such as 3D telepresence, game modeling, metaverse [40], etc. Early works show promise under equipment-assisted settings, requiring expensive dense camera rigs [12] and tedious calibration procedures [15]. Parametric models, such as SMPL [36] and SMPL-X [47], have been introduced to model a naked human body with constrained parameters. With the witnessed success of deep learning in many vision tasks, many deep learning methods have been proposed to regress the parameters of such parametric human models from a single [16, 66, 67] or multiple images [6, 13, 68]. However, these methods can only reconstruct a minimally-clothed human model without many details (*e.g.*, hairs and clothes).

Recently, methods based on implicit shape representation have reported encouraging performance, showing promising reconstruction with increased details in both single-view [23, 26, 52, 63, 71] and multi-view [24, 25, 54, 55] settings.

Despite the stunning results reported by the above methods, their reconstructions remain far from perfect, restricting their practical applications. For instance, state-of-the-art (SOTA) single-view methods, such as PIFuHD [53], PaMIR [71], and ARCH++ [23], struggle with many self-occluding non-frontal human poses that widely occur in the real world. ICON [63] can handle these cases but the reconstructions contain serious artifacts (see Fig. 3). On the other hand, many multi-view methods depend on calibrated cameras (*e.g.*, [24, 54, 55]) which are tedious to obtain in practice. Hence, how to carry out multi-view reconstruction with uncalibrated cameras is an important topic to study. Meanwhile, effective multi-view feature fusion is another key factor for robust mutli-view reconstruction. Prior techniques for multi-view feature fusion include average pooling [52], SMPL-visibility [63], and attention-based mechanism [70]. However, the reconstruction results based on these fusion techniques still contain notable artifacts in many cases. This indicates that more efforts are needed to derive a better multi-view feature fusion for more robust reconstruction.

In this paper, to extract more clothed human details flexibly and robustly from a single RGB image or uncalibrated multi-view RGB images, we present a novel framework, named SeSDF, that employs the parametric model SMPL-X [47] as a 3D prior and combines the merits of both implicit and explicit representations. SeSDF takes a single image or multi-view images as input to predict the occupancy for each 3D location in the space representing the human model. To reconstruct high-frequency details such as hairs and clothes, we introduce a self-evolved signed distance field module that learns to deform the signed distance field (SDF) derived from the fitted SMPL-X model using the input images. The resulting SDF can reflect more accurate geometry details than SMPL-X, which inherently deviates from the actual clothed human model. The SDF refined by our SeSDF module is further encoded to allow for 3D reconstruction with better geometry details.

Besides reconstructing a faithful 3D clothed human model from a single image, our SeSDF framework can also work with uncalibrated multi-view images to generate clothed human avatar with enhanced appearance (see Fig. 1). To this end, we first propose a self-calibration method by fitting a shared SMPL-X model across multi-view images and projecting the shared model to different images based on the optimized rigid body motion for each input image. Further, we propose an occlusion-aware feature fusing strategy by probing the visibility of each 3D point under different views through ray-tracing, leveraging the SMPL-X model, such that features from visible views will contribute more to the fused feature while those from invisible views will be suppressed.

The contributions are summarized as follows:

- We propose a flexible framework that, by leveraging the SMPL-X model as a shape prior, can take an arbitrary number of uncalibrated images to perform high-fidelity clothed human reconstruction. At the core of our framework is a self-evolved signed distance field (SeSDF) module for recovering faithful geometry details.

- For uncalibrated multi-view reconstruction, we propose a simple self-calibration method through leveraging the SMPL-X model, as well as an occlusion-aware feature fusion strategy which takes into account the visibility of a space point under different views via ray-tracing.

- We thoroughly evaluate our framework on public benchmarks. SeSDF exhibits superior performances over current state-of-the-arts both qualitatively and quantitatively.

2. Related Work

Parametric 3D human model Parametric 3D human reconstruction [21, 30, 36, 44, 47] paves the path for 3D human model. SMPL [36, 67] was built on convolutional networks and statistics collected from a wide range of human populations. SMPL-X [47, 66] extended SMPL with more body joints, keypoints, and expression parameters. Later substantial supervision strategies were proposed to regress the model parameters. Examples include semantic segmentation [43, 64], human body silhouette [49], joints [16], motion dynamics [30, 33], texture consistency [48], and attention-regressor [31]. Multi-view [6, 13, 68, 69] and multi-person [57, 58] setups were also sought for various scenarios. Complex human clothing and hair topology were also the goals for later parametric methods [3, 4, 29, 62, 73]. However, due to the minimal-clothing assumption of the parametric models, these methods can retain little geometry details such as hairs and clothes in their reconstructions.

Single-view (implicit) human reconstruction Many methods have been proposed in the deep learning era to reconstruct a 3D human model from a single image. For example, voxel-based methods [18, 27, 59, 60, 72, 72] reconstruct a volumetric representation of the human model. Two-stage inference [56], visual hull [41] and front-back depth method [17] were then proposed to improve the reconstruction performance and efficiency. However, the high computational cost of these methods limits their real-world applications. Implicit function methods have recently been introduced to predict either the occupancy field [10, 38] or signed distance field [8, 28, 45] using neural networks. PIFu [52] presents promising results with a pixel-aligned implicit function. Follow-up works exploited 3D features

to solve the depth-ambiguity problem [71], achieve animatable model [23, 26], or improve the reconstruction details [7]. PIFuHD [53] and FITE [35] apply a multi-stage method to improve clothing topology in the reconstruction. Geo-PIFu [22] proposes to compute 3D features directly from images. ICON [63] exploits the SMPL-based signed distance filed as guidance to help produce a robust reconstruction, with a smpl-visibility based method to choose features from front or back sides. PHORHUM [5] explores shading and illumination information to infer better geometry and texture. Encouraging improvements have been reported by these methods. However, their performance on non-frontal human images are still far from being robust and satisfactory.

Multi-view implicit human reconstruction Multi-view reconstruction is conceptually a more natural choice as information from different views can compensate each other for better reconstruction. Under a calibrated setting, PIFu [52] proposes to fuse features from multi-views by average pooling after being embedded using an MLP. StereoPIFu [24] takes as input a pair of stereo images, and performs depth-aware reconstruction. DiffuStereo [55] employs diffusion models for multi-view stereo human reconstruction. DoubleField [54] combines the surface field and radiance fields to improve both geometry and texture inference. Effi-MVS [61] explores the dynamic cost volume for reduced computational cost. DeepMultiCap [70] handles the multi-person reconstruction with multi-view images. Neural-Human-Performer [32] proposes a multi-view transformer to fuse features for multi-view reconstruction. Despite promising results have been achieved, these methods require a large number of views and camera calibration. MVP-Human [74] makes use of the deformation fields and handles multi-view inputs under an uncalibrated setting, but the reconstruction quality lags far behind the calibrated methods. For calibrated methods, their models still produce undesirable artifacts for many self-occluding non-frontal poses and miss many geometry details.

3. Method

We propose an implicit function based framework for reconstructing 3D clothed human models from a single image or uncalibrated multi-view images (see Fig. 2). At the core of our framework are a self-evolved signed distance field (SeSDF) module and an occlusion-aware feature fusion module. Next, we will first introduce the preliminary for SeSDF in Sec. 3.1. We will then describe SeSDF for single-view reconstruction in Sec. 3.2, followed by uncalibrated multi-view reconstruction in Sec. 3.3.

3.1. Preliminary

Surface by implicit function Implicit functions [14] have been shown to be effective for 3D reconstruction from im-

ages or videos. Typically, an implicit function is implemented as an MLP, which takes the 3D coordinate X and an optionally associated feature F of a 3D point as input, and outputs the corresponding occupancy as a scalar value. Commonly employed features include pixel-aligned image features [52, 53] and space-aligned 3D features [7, 26, 71]. Formally, the implicit function can be written as

$$f_v(X, F) \mapsto [0, 1] \in \mathbb{R}. \tag{1}$$

SMPL-X The parametric human model SMPL-X represents the 3D shape by incorporating body vertices, joints, face and hands landmarks, and expression parameters. Formally, by assembling pose parameters θ, shape parameters β, and facial expression parameters ϕ, SMPL-X human model can be expressed by a function $M(\cdot)$ as

$$
\begin{aligned}
M(\beta, \theta, \phi) &= W(T_p(\beta, \theta, \phi), J(\beta), \theta, \mathcal{W}), \\
T_p(\beta, \theta, \phi) &= \bar{T} + B_S(\beta; \mathcal{S}) + B_E(\phi; \mathcal{E}) + B_p(\theta; \mathcal{P}),
\end{aligned}
$$

where $W(\cdot)$ is the linear blend skinning function, $J(\beta)$ is the joint location, and $\mathcal{W}$ denotes the blend weights. $T_p(\cdot)$ poses the SMPL-X canonical model $\bar{T}$ by shape blend shape function B_S, expression blend shape function B_E, and pose blend shape function B_P. $\mathcal{S}$ and $\mathcal{P}$ are the principal components of vertex displacements, and $\mathcal{E}$ represents the principal components capturing variations of facial expressions.

SMPL-X optimization Given a human image, we can fit the SMPL-X model parameters by [16]. However, such fitted models often suffer from misalignment. We use the well-aligned SMPL-X models provided in the training data to provide shape prior to train our framework. At test time, we do not have such well-aligned models. Hence, to avoid errors and artifacts caused by the misalignment during testing, we optimize the SMPL-X parameters by projecting the SMPL-X model onto the input image and maximizing the Intersection over Union (IoU) between the projection and body mask [9, 63], and minimizing the 2D keypoint distances. Please refer to the supplementary for details.

3.2. Single-view reconstruction

Given an input image $\boldsymbol{I}$ and an optimized SMPL-X mesh $\boldsymbol{S}$, SeSDF incorporates pixel-aligned image features F_{2D} extracted from $\boldsymbol{I}$, space-aligned 3D features F_{3D} extracted from $\boldsymbol{S}$, and the self-evolved signed distance field d to predict occupancy using an MLP. By applying marching cube [34] on the occupancy predictions in the 3D space, SeSDF robustly reconstructs a 3D clothed human model with faithful details, including the clothing topology. Given a 3D point X, the implicit function for single-view reconstruction can be formulated as

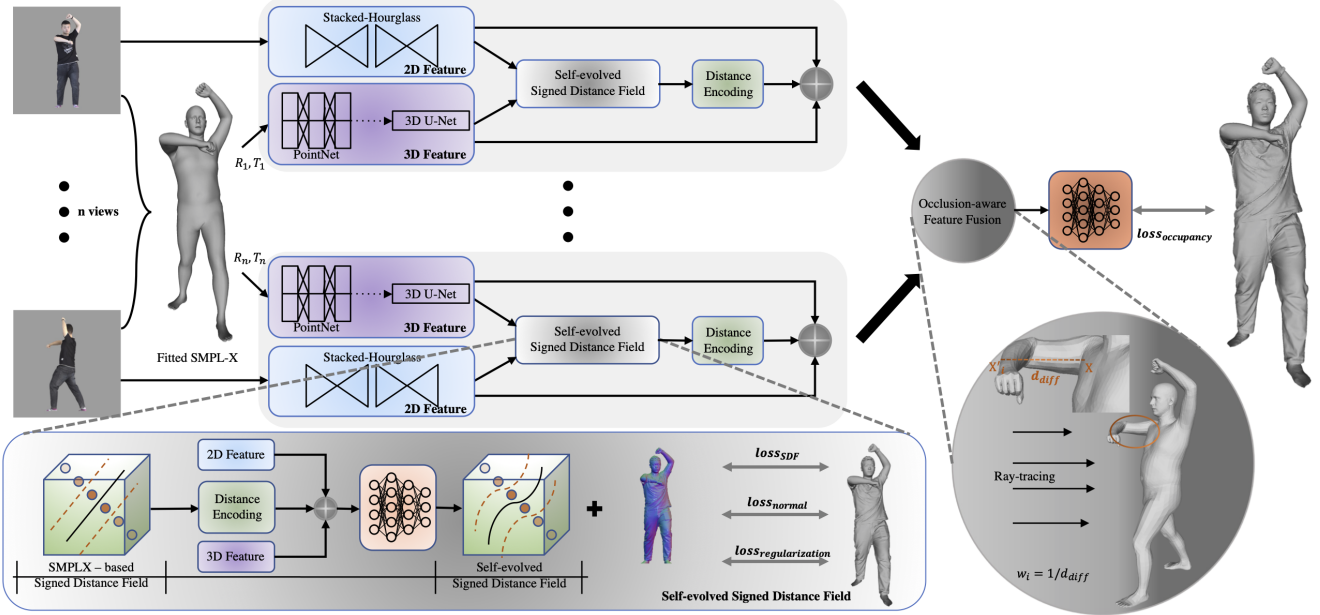$$f_o(F_{2D}(X), F_{3D}(X), \mathcal{D}(d(X)), \mathbf{n}(X), Z(X)) \mapsto [0, 1], \tag{2}$$

Figure 2. Overview of our proposed SeSDF framework. It takes a single image or uncalibrated multi-view images as input to predict the occupancy for each 3D location in the space for shape reconstruction. We first fit a SMPL-X model to the input image(s) which serves as a shape prior and provides space-aligned 3D features. At the core of our method is our proposed self-evolved signed distance field (SeSDF) module, which learns a refined signed distance field (SDF) to better reflect the actual clothed human subject rather than the minimally-clothed SMPL-X model. SeSDF module takes pixel-aligned image features, space-aligned 3D features, SDF with distance encoding, and SMPL-X vertex normal as input to predict the refined SDF and normal value. For multi-view reconstruction, we propose a SMPL-X based self-calibration method to eliminate the need for manual calibration and an occlusion-aware feature fusion strategy to effectively accumulate features from different views.

where $\mathcal{D}(\cdot)$ and $Z(\cdot)$ denote distance encoding and depth value respectively, and $d(\cdot)$ and $\mathbf{n}(\cdot)$ are the signed distance and normal output by our SeSDF module.

Pixel-aligned image feature We apply a stacked-hourglass encoder [42, 52] to the input image $I \in \mathbb{R}^{512 \times 512 \times 3}$ to produce an image feature map $E_I(I) \in \mathbb{R}^{256 \times 128 \times 128}$. We then obtain the pixel-aligned image feature for the projection of a 3D point using bi-linear interpolation.

Space-aligned 3D feature We apply PointNet [51] and 3D-UNet [11] to extract 3D features [7, 50] from the optimized SMPL-X mesh. Our 3D encoder first processes the whole 10,475 vertices of the mesh to generate hierarchical 3D point features, which are transformed into a 3D feature volume via average pooling. 3D-UNet is then applied to output the final 3D feature volume $E_S(\boldsymbol{S}) \in \mathbb{R}^{128 \times 64 \times 64 \times 64}$. Finally we obtain the space-aligned 3D feature for a 3D point via tri-linear interpolation.

Self-evolved signed distance field The signed distance field (SDF) derived from the SMPL-X model [63] provides a strong prior for estimating the occupancy. However, the SMPL-X model does not model geometric details of the human, such as the cloth topology, and thus the resulting SDF will unavoidably deviate from the true SDF of the actual clothed human model. Directly employing this SDF

for occupancy prediction will therefore induce error in the reconstruction. To remedy this problem and recover plausible geometry details in the reconstruction, we propose a novel self-evolved signed distance filed (SeSDF) module. This module is trained to evolve the SDF derived from the SMPL-X model to better reflect the actual clothed human model. Instead of predicting only the signed distance $d(\cdot) \in \mathbb{R}$, our SeSDF module further predict the normal $\mathbf{n}(\cdot) \in \mathbb{R}^3$ which provides additional surface orientation information to guide the implicit function to better predict occupancy and capture fine shape details. Concretely, SeSDF module is implemented as an MLP and formulated as

$$f_{sd}(F_{2D}(X), F_{3D}(X), \mathcal{D}(d'(X)), \mathbf{n}'(X)) \mapsto (d, \mathbf{n}), \quad (3)$$

where $d'(\cdot) \in \mathbb{R}$ denotes the SDF derived from the SMPL-X model and $\mathbf{n}'(\cdot) \in \mathbb{R}^3$ represents the normal obtained from the closest SMPL-X vertex to X.

Distance encoding Inspired by the positional encoding adopted in NeRF [39], we adopt a distance encoding that maps the signed distance d to a higher dimensional space using high frequency functions:

$$\mathcal{D}(d) = (d, \sin(2^0 \pi d), \cos(2^0 \pi d), ..., \sin(2^L \pi d), \cos(2^L \pi d)). \quad (4)$$

In our experiments, we set $L = 5$. The encoded signed distance $\mathcal{D}(d)$ is concatenated with the predicted normal $\mathbf{n}(\cdot)$,

2D feature $F_{2D}(\cdot)$ and 3D feature $F_{3D}(\cdot)$, and fed to the implicit function $f_o(\cdot)$ for occupancy prediction.

Training objectives To consistently regress SeSDF and simulate realistic cloth topology, we consider two groups of sampling points, namely surface samples G_s, and occupancy-target samples G_o. Points in G_s are uniformly sampled on the mesh surface, while points in G_o are sampled following the strategy proposed in PIFu [52].

For points in G_s, we adopt the following loss function

$$\mathcal{L}_s = \frac{1}{|G_s|} \sum_{X \in G_s} \lambda_d |d(X)| + \lambda_n ||\mathbf{n}(X) - \mathbf{n}_{gt}(X)||, \quad (5)$$

where $\mathbf{n}_{gt}(X)$ denotes the ground-truth normal at the surface point X. For points in G_o, we adopt the following loss function

$$\mathcal{L}_o = \frac{1}{|G_o|} \sum_{X \in G_o} BCE(\mathcal{O}(X) - \mathcal{O}_{gt}(X)), \quad (6)$$

where $\mathcal{O}(X)$ denotes the predicted occupancy at X computed using Eq. (3) and Eq. (2), $\mathcal{O}_{gt}(X)$ is the ground-truth occupancy at X, and $BCE(\cdot)$ denotes the binary cross entropy. Following IGR [20], we add an eikonal normal regularization term to smooth the surface by:

$$\mathcal{L}_r = \frac{1}{|G_o|} \sum_{X \in G_o} (||\mathbf{n}(X)|| - 1)^2. \quad (7)$$

The overall loss function is given by

$$\mathcal{L}_{SeSDF} = \lambda_s \mathcal{L}_s + \lambda_o \mathcal{L}_o + \lambda_r \mathcal{L}_r, \quad (8)$$

where λ_s, λ_o, and λ_r are weights to balance the effect of the respective loss.

3.3. Uncalibrated multi-view reconstruction

To allow our SeSDF framework work with uncalibrated multi-view images, we present a simple self-calibration method to estimate the relative camera pose for each image through fitting a SMPL-X model to all the input images. In addition, we propose an occlusion-aware feature fusion strategy to fuse features from different views effectively. Ideally, we would like to fuse only features from non-occluded views as features from occluded views are just noise which at best provide zero contribution to occupancy prediction but may also seriously degrade the fused feature in the worse case.

Our multi-view framework is illustrated in Fig. 2. In principle, it can work with an arbitrary number of views and the number of views used in training and testing do not even need to be the same.

Self-calibration via SMPL-X model fitting Given n uncalibrated multi-view images, we first fit a SMPL-X model to each image independently using the method described in

Sec. 3.1. We then take the averages of the shape, pose and expression parameters of these n fitted models and use them to initialize a shared SMPL-X model. Next, we optimize the shape, pose and expression parameters of this shared SMPL-X model and the global orientation parameters for the n views simultaneously by maximizing the IoU between the projection and body mask, and minimizing the 2d keypoint distances on each of the n images. After optimization, the global orientation parameters for the n views then give us the rigid body motion that transforms a point from the SMPL-X model centered coordinate system to the camera centered coordinate system of each view.

To perform multi-view reconstruction, we can sample 3D points in the model centered coordinate system, transform them into the respective camera coordinate systems, and compute 2D/3D features and the self-evolved signed distance from each view. Finally, these features from n different views are fused by our occlusion-aware feature fusion strategy (to be introduced next) and fed to the implicit function $f_o(\cdot)$ to predict the occupancy value.

Occlusion-aware feature fusion Conceptually, features obtained from different views would give rise to predictions of different quality. For instance, image feature for a 3D point extracted from a frontal non-occluded view would be expected to give the best prediction, whereas image feature for the same point but extracted from an occluded view or a lateral view would be expected to give poor prediction. Hence, average pooling would in principle not be the best way for feature fusion. Researchers therefore explored different fusion strategies, such as SMPL-visibility based [63] and attention based [70] approaches, but all of these still produce undesirable artifacts in their reconstructions (see Fig. 7). To lessen the effect of features extracted from non-optimal views, we propose a novel occlusion-aware feature fusion strategy based on ray tracing [19] with the fitted SMPL-X model. Given a 3D point X, we cast a ray from X to the i-th view through orthogonal projection and find the intersection point X'_i with the SMPL-X model that is the closest to the image (see Fig. 2). We then compute a weight for the feature $F_i(X)$ extracted from the i-th view by

$$w_i(X) = \frac{1}{|Z_i(X) - Z_i(X'_i)|}. \quad (9)$$

Finally, we fuse all features from the n views by

$$F_{fused}(X) = \frac{\sum_{i=1}^n w_i(X) F_i(X)}{\sum_{i=1}^n w_i(X)}. \quad (10)$$

4. Experiments

We evaluate the performance of our SeSDF both qualitatively and quantitatively, and provide comparisons with other SOTA methods under both single-view and multi-view setups.

Table 1. Qualitative comparison on the single-view setting. We report Chamfer Distance and P2S metrics (numbers the lower the better).

| Method | Feature Included | | | | | Quantitative Number | | | |
| | | | | | | THuman2.0 | | CAPE | |
	2D Feature	Normal Information	3D Feature	SMPLX Signed Distance	self-evolved Signed Distance	Chamfer Distance↓	P2S↓	Chmafer Distance↓	P2S↓
PIFu [52]	✓					2.403	2.388	3.059	2.818
PIFuHD [53]	✓	✓				2.008	1.965	2.571	2.427
PaMIR [71]	✓		✓			1.478	1.451	1.682	1.438
ICON [63]	✓	✓		✓		1.301	1.259	1.533	1.431
Ours w/o DE	✓	✓	✓		✓	1.099	1.034	1.377	1.261
Ours w/o SeSDF	✓	✓	✓	✓		1.245	1.200	1.473	1.390
Ours	✓	✓	✓		✓	**1.027**	**0.971**	**1.303**	**1.219**

Table 2. Quantitative comparison on the multi-view setting.

| Method | THuman2.0 | | CAPE | |
	Chamfer Distance↓	P2S↓	Chmafer Distance↓	P2S↓
multi-view PIFu [52]	0.864	0.838	0.920	0.876
multi-view PIFuHD [53]	0.831	0.804	0.883	0.849
multi-view PaMIR [71]	0.807	0.762	0.858	0.819
Ours w/o DE	0.795	0.749	0.849	0.811
Ours w/o SeSDF	0.802	0.755	0.854	0.815
Ours	**0.791**	**0.742**	**0.841**	**0.800**
Ours w/ Attention	0.794	0.743	0.845	0.803
Ours w/ SMPLX-Vis	0.841	0.832	0.925	0.917
Ours w/ Normal-Fusion	0.801	0.754	0.850	0.812

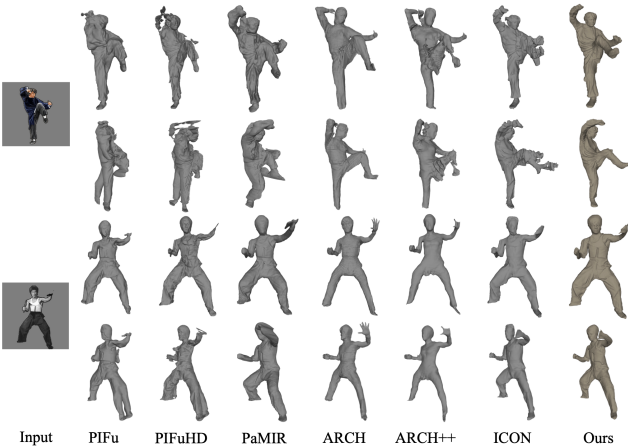Input   PIFu   PIFuHD   PaMIR   ARCH   ARCH++   ICON   Ours

Figure 3. Qualitative comparison with SOTA methods on real-world images. Our method can reconstruct fine details and work robustly under challenge poses. Best viewed in PDF with zoom.

4.1. Implementation details

For each training subject, we sample (a) 5,000 surface points as G_s, which are evenly distributed on the mesh surface; (b) 5,000 occupancy-target points as G_o in the 3D space, for which we follow the sampling strategy in [52]. In particular, 15/16 points in G_o are evenly sampled on the mesh surface. Gaussian perturbation is then applied to them along the surface normal direction. The rest 1/16 points are randomly sampled within the predefined 3D space where the mesh lies in.

We employ Rembg [1] to segment out the background for real-world images and use Kaolin [2] to calculate the SDF from the SMPL-X model. We apply PIXIE [16] to estimate the SMPL-X model for the segmented human subject. To compute the image features, we implement a stacked-hourglass image encoder with bi-linear interpolation to extract pixel-aligned image feature $F_{2D}(\cdot) \in \mathbb{R}^{256}$. We follow [7, 50] to employ PointNet with 3D-UNet and tri-linear interpolation to extract space-aligned 3D features $F_{3D}(\cdot) \in \mathbb{R}^{128}$. Distance encoding maps the signed distance to a higher dimensional space $\mathcal{D}(d) \in \mathbb{R}^{13}$ using the high-frequency functions as defined in Eq. (4). We implement our framework in PyTorch [46]. Our SeSDF framework is trained separately for 12 epochs with the learning rate starting from 1e-4 and updated by the factor 0.1 for every 4 epochs, for both single-view and multi-view experiments. Code for SeSDF will be made publicly available. Project page: https://yukangcao.github.io/SeSDF/.

4.2. Datasets and evaluation metrics

We employ the THUman2.0 dataset [65] as our primary testbed. The dataset comprises 526 high-fidelity meshes captured via dense DSLR rigs. Each subject within the dataset is accompanied by its corresponding texture map and fitted SMPL-X parameters. We split the THUman2.0 dataset into a training set with 465 subjects, and a test set with 61 subjects. We render each human subject into 360 images, sequentially separated by $1°$ around the subject, with different albedo to augment the dataset to carry out the experiments.

To further show the generalizability of SeSDF, we also compare our method with others on the CAPE [37] test set, which contains 150 subjects. To conduct quantitative evaluations of different methods, we employ standard Point-to-Surface distance (P2S) and Chamfer distance metrics as evaluation criteria. We compare the reconstructed models from images captured at different angles, *i.e.*, $0°$, $45°$, $90°$, $135°$, $180°$, $225°$, $270°$, and $315°$, against the ground truth under corresponding orientations. Besides, we use real-world images for extra qualitative evaluation.

4.3. Comparison on single-view setting

We first evaluate and compare the performance of single-view 3D human reconstruction between ours and SOTA methods. Specifically, we compare with PIFu [52], PI-FuHD [53], PaMIR [71], ARCH [26], ARCH++ [23] and ICON [63]. PHORHUM [5] is not compared here due to the absence of publicly available code base. We retrain PIFu, PIFuHD, PaMIR, and ICON on the THUman2.0 dataset for
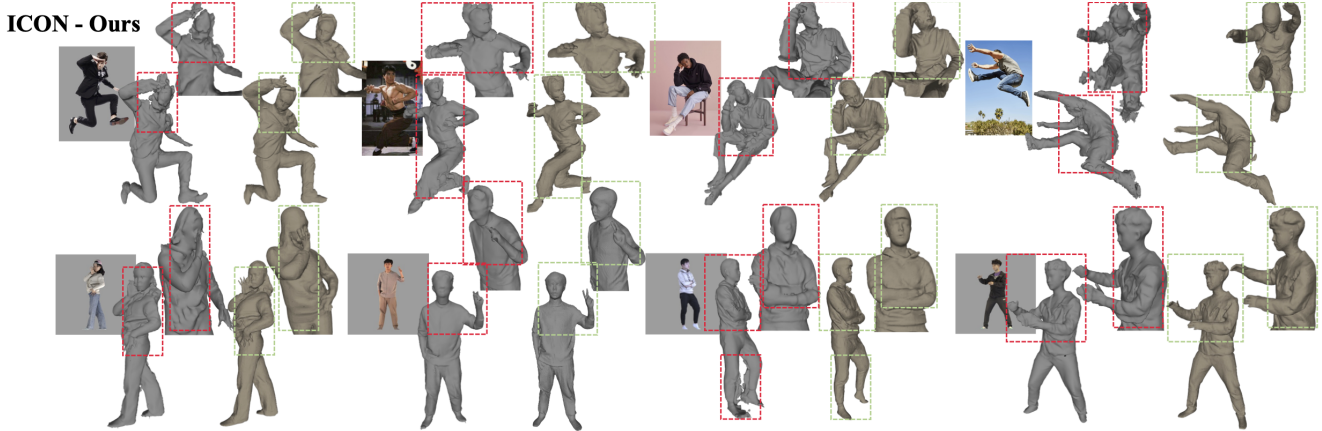
Figure 4. Comparison with ICON on real-world (top) and THUman2.0 data (bottom). Ours (SeSDF) can produce results with more fine geometry details. Best viewed in PDF with zoom.

a fair comparison, and obtain the reconstructions of ARCH and ARCH++ with the help of their authors.

Quantitative comparison and ablation In Table 1, we compare our method with the SOTA methods, *i.e.*, PIFu, PIFuHD, PaMIR, and ICON on the THUman2.0 and CAPE datasets. As can be seen, for all metrics, our method significantly outperforms the SOTA by a large margin on both datasets. Among these SOTA methods, ICON also employs the SMPL-X model to extract SDF to guide learning, while our method consistently outperforms ICON across the board. We also validate the effectiveness of different components of our method. All components have positive effects on the reconstruction results, while the SeSDF module plays the most crucial role in the improvement. By comparing the last two rows, we can see that the SeSDF module brings much more improvement than the SDF derived from the SMPL-X model. By comparing the last row with all others, we can observe that our SeSDF brings the most improvement among all components. By comparing "Ours" and "Ours w/o DE", we can see that our proposed distance encoding is also effective in improving the performance.

Qualitative comparison We present qualitative comparison in Fig. 3 on real-world images with SOTA methods. Our proposed method has the capability to reconstruct fine geometry details and handle the inputs with complicated poses. Among the SOTA methods, ICON also employs SMPL-X to improve the robustness, though in a different way from our work. We provide a further comparison with ICON in Fig. 4 on THUman2.0 and real-world images. As can be seen, our method notably outperforms ICON with more fine details in all cases.

4.4. Comparison on multi-view setting

For multi-view reconstruction, we compare with the SOTA methods PIFu, PIFuHD, and PaMIR which allow multi-view reconstruction. In the experiments, we use three input views unless stated otherwise.

Quantitative comparison and ablation We report the comparison results in Table 2. To have a fair comparison with other methods which are calibrated, we also implemented our self-calibration method for all of them. We can see that our method outperforms other methods in all cases. Similar to the single-view experiments, by comparing "Ours" against "Ours /wo DE" and "Ours /wo SeSDF", we can see that our *distance encoding* and *SeSDF* play critical roles in the performance.

Qualitative comparison We present qualitative comparison in Fig. 5 for multi-view reconstruction. The reconstruction results of our method retain more details, *e.g.*, clothing wrinkles and hair topology, which are contained in the ground truth. We provide more results with different numbers of views in the supplementary.

4.4.1 Further analysis

Self-calibration [74] also introduced an alternative method for self-calibration of multi-view reconstruction by introducing a deformation field to deform points from the canonical space to the camera space, through SMPL linear blend skinning (LBS) weights. We also follow the paper to implement it for SMPL-X in our framework and compare it with our method in Fig. 6. We observe wide-spreading artifacts across the whole body of the reconstruction using the deformation field for the uncalibrated setting. The deformation field is error-prone because the LBS weights are computed to deform the parametric human model, which inherently lacks the fine details of the actual human subject, limiting its final performance and introducing the artifacts for reconstruction.

Feature fusion We compare our occlusion-aware feature fusion strategy with other alternatives, namely (1) AVG-Pool [52], (2) SMPLX-visibility [63], and (3) Attention-based method [32] in Fig. 7. As can be seen, our occlusion-aware strategy achieves better performance than the others. See supplementary for more analysis.
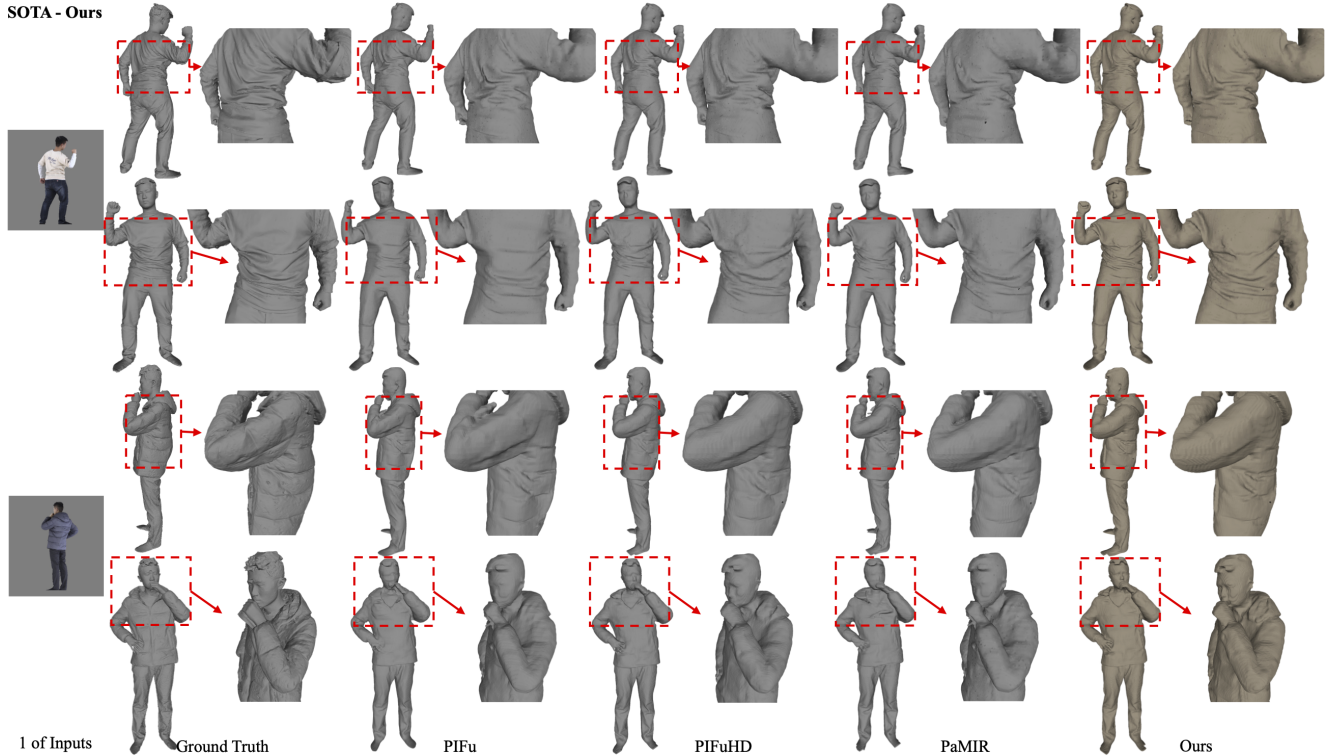
Figure 5. Comparison with SOTA methods on the multi-view setting. Ours (SeSDF) consistently retains more geometry details, *e.g.*, clothing wrinkles, hairs, and faces.
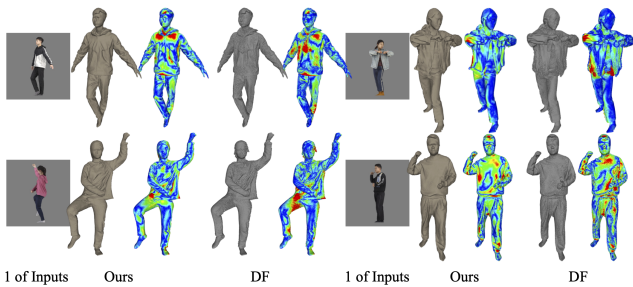

Figure 6. Comparing our self-calibration method ("Ours") with an SMPL-X deformation field alternative ("DF"). Best viewed in PDF with zoom.

Figure 7. Comparing our occlusion-aware feature fusion with other alternative fusion strategies.

5. Conclusion

In this paper, we have presented a new framework for single- and multi-view clothed human reconstruction under an uncalibrated setting. Our framework can take an arbitrary number of input images to reconstruct faithful 3D human models. We have proposed a self-evolved signed distance (SeSDF) module to recover more geometry details, a method for self-calibrating multi-view images via SMPL-X model fitting, and an effective multi-view occlusion-aware feature fusion strategy to aggregate useful features across views for robust reconstruction. Our method notably outperforms previous SOTA methods in both terms of qualitative and quantitative results.

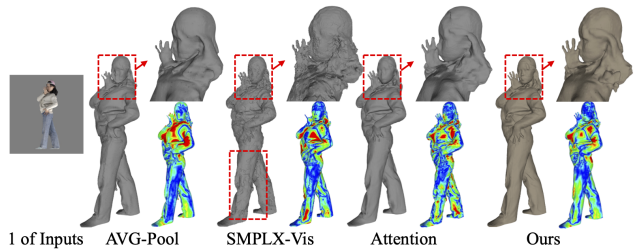**Limitation** Though our SeSDF can deal with various poses and retain fine clothing topology, it still has difficulties when handling extremely loose clothes, *e.g.*, dresses, which deviate significantly from the parametric prior.

Potential Social Impact Our method takes a step towards practical tools for real-world application, as it does not require expensive equipment to get the 3D models. However, despite its intended positive usage, there might be potential risks of falsifying human avatars, which unfortunately compromise personal privacy. Transparency and authentication will be helpful to prevent these vicious uses.

References

[1] https://github.com/danielgatis/rembg. 6

[2] https : / / github . com / NVIDIAGameWorks / kaolin. 6

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[5] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3, 6

[6] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies. *arXiv preprint arXiv:2104.03978*, 2021. 1, 2

[7] Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K Wong. Jiff: Jointly-aligned implicit face function for high quality single view clothed human reconstruction. In *IEEE conference on Computer Vision and Pattern Recognition*, 2022. 3, 4, 6

[8] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *European Conference on Computer Vision*, 2020. 2

[9] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *arXiv preprint arXiv:2204.11798*, 2022. 3

[10] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[11] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, 2016. 4

[12] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam G. Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 2015. 1

[13] Zijian Dong, Jie Song, Xu Chen, Chen Guo, and Otmar Hilliges. Shape-aware multi-person pose estimation from multi-view images. In *IEEE International Conference on Computer Vision*, 2021. 1, 2

[14] Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings*. 2009. 3

[15] Massimiliano Favalli, Alessandro Fornaciai, Ilaria Isola, Simone Tarquini, and Luca Nannipieri. Multiview 3d reconstruction in geosciences. *Computers Geosciences*, 2012. 1

[16] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision*, 2021. 1, 2, 3, 6

[17] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *IEEE International Conference on Computer Vision*, 2019. 2

[18] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In *European Conference on Computer Vision*, 2018. 2

[19] Andrew S Glassner. *An introduction to ray tracing*. Morgan Kaufmann, 1989. 5

[20] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *International Conference on Machine Learning*, 2020. 5

[21] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *IEEE International Conference on Computer Vision*, 2009. 2

[22] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Conference on Neural Information Processing Systems*, 2020. 3

[23] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *IEEE International Conference on Computer Vision*, 2021. 2, 3, 6

[24] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *IEEE conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[25] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *European Conference on Computer Vision*, 2018. 2

[26] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3, 6

[27] Aaron S Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *European Conference on Computer Vision Workshops*, 2018. 2

[28] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[29] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE conference on computer vision and pattern recognition*, 2018. 2

[30] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[31] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D

human body estimation. In *IEEE International Conference on Computer Vision*, 2021. 2

[32] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Conference on Neural Information Processing Systems*, 2021. 3, 7

[33] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[34] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics*, 2018. 3

[35] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *European Conference on Computer Vision*, 2022. 3

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015. 1, 2

[37] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[38] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2

[39] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 4

[40] Stylianos Mystakidis. Metaverse. *Encyclopedia*, 2022. 1

[41] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[42] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, 2016. 4

[43] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision*, 2018. 2

[44] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *European Conference on Computer Vision*, 2020. 2

[45] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[46] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration, 2017. 6

[47] 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 *IEEE conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[48] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *IEEE International Conference on Computer Vision*, 2019. 2

[49] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[50] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, 2020. 4, 6

[51] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE conference on Computer Vision and Pattern Recognition*, 2017. 4

[52] 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 *IEEE International Conference on Computer Vision*, 2019. 2, 3, 4, 5, 6, 7

[53] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3, 6

[54] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *IEEE conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[55] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *European Conference on Computer Vision*, 2022. 2, 3

[56] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *IEEE International Conference on Computer Vision*, 2019. 2

[57] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *IEEE International Conference on Computer Vision*, 2021. 2

[58] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[59] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *European Conference on Computer Vision*, 2018. 2

[60] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *European Conference on Computer Vision*, 2018. 2

[61] Shaoqian Wang, Bo Li, and Yuchao Dai. Efficient multi-view stereo by iterative dynamic cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[62] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[63] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 4, 5, 6, 7

[64] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *IEEE International Conference on Computer Vision*, 2019. 2

[65] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 6

[66] 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. *arXiv preprint arXiv:2207.06400*, 2022. 1, 2

[67] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *IEEE International Conference on Computer Vision*, 2021. 1, 2

[68] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 2021. 1, 2

[69] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[70] 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. *arXiv preprint arXiv:2105.00261*, 2021. 2, 3, 5

[71] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3, 6

[72] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *IEEE International Conference on Computer Vision*, 2019. 2

[73] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[74] Xiangyu Zhu, Tingting Liao, Jiangjing Lyu, Xiang Yan, Yunfeng Wang, Kan Guo, Qiong Cao, Stan Z Li, and Zhen Lei. Mvp-human dataset for 3d human avatar reconstruction from unconstrained frames. *arXiv preprint arXiv:2204.11184*, 2022. 3, 7