REC-MV: REconstructing 3D Dynamic Cloth from Monocular Videos

Lingteng Qiu[1]* Guanying Chen[1,2]* Jiapeng Zhou[1] Mutian Xu[1]
Junle Wang[3] Xiaoguang Han[1,2]†

[1]SSE, CUHKSZ [2]FNii, CUHKSZ [3]Tencent

Abstract

Reconstructing dynamic 3D garment surfaces with open boundaries from monocular videos is an important problem as it provides a practical and low-cost solution for clothes digitization. Recent neural rendering methods achieve high-quality dynamic clothed human reconstruction results from monocular video, but these methods cannot separate the garment surface from the body. Moreover, despite existing garment reconstruction methods based on feature curve representation demonstrating impressive results for garment reconstruction from a single image, they struggle to generate temporally consistent surfaces for the video input. To address the above limitations, in this paper, we formulate this task as an optimization problem of 3D garment feature curves and surface reconstruction from monocular video. We introduce a novel approach, called **REC-MV***, to jointly optimize the explicit feature curves and the implicit signed distance field (SDF) of the garments. Then the open garment meshes can be extracted via garment template registration in the canonical space. Experiments on multiple casually captured datasets show that our approach outperforms existing methods and can produce high-quality dynamic garment surfaces. The source code is available at https://github.com/GAP-LAB-CUHK-SZ/REC-MV.*

1. Introduction

High-fidelity clothes digitization plays an essential role in various human-related vision applications such as virtual shopping, film, and gaming. In our daily life, humans are always in a moving status, driving their clothes to move together. To realize this very common scenario, it is indispensable to gain dynamic garments in real applications. Thanks to the rapid development of mobile devices in terms of digital cameras, processors, and storage, shooting a monocular video in the wild becomes highly convenient and accessible for general customers. In this paper, our goal is definite – *extracting* ***dynamic 3D garments*** *from* ***monocular videos***, which is significantly meaningful and valuable for practical applications, but is yet an uncultivated land with many challenges.

Figure 1. *Can we extract* ***dynamic 3D garments*** *from* ***monocular videos***? The answer is Yes! By jointly optimizing the dynamic feature curves and garment surface followed by non-rigid template registration, our method can reconstruct high-fidelity and temporally consistent garment meshes with open boundaries.

We attempt to seek a new solution to this open problem and start by revisiting existing works from two mainstreams. i) Leveraging the success of neural rendering methods [35, 37, 57], several works are able to reconstruct dynamic clothed humans from monocular videos [8, 19, 29, 47, 49], by representing the body surface with an implicit function in the canonical space and apply skinning based deformation for motion modeling. One naive way to achieve our goal is: first to get the clothed human through these methods and separate the garments from human bodies. However, such a separation job requires laborious and non-trivial processing by professional artists, which is neither straightforward nor feasible for general application scenarios. ii) As for garment reconstruction, many methods [5, 10, 20, 61, 62] make it possible to reconstruct high-quality garment meshes from single-view images in the wild. Specifically, ReEF [62] estimates 3D fea-

* Equal contribution.
† Corresponding author: hanxiaoguang@cuhk.edu.cn.

ture curves* and an implicit surface field [34] for non-rigid garment template registration. Nonetheless, these methods struggle to produce temporally consistent surfaces when taking videos as inputs.

The above discussion motivates us to combine the merits of both the dynamic surface modeling in recent neural rendering methods and the explicit curve representation for garment modeling. To this end, we try to delineate a new path towards our goal: *optimizing dynamic explicit feature curves and implicit garment surface from monocular videos*, to extract temporally consistent garment meshes with open boundaries. We represent the explicit curves and implicit surface in the canonical space with skinning-based motion modeling, and optimize them by 2D supervision automatically extracted from the video (*e.g.*, image intensities, garment masks, and visible feature curves). After that, the open garment meshes can be extracted by a garment template registration in the canonical space (see Fig. 1).

We strive to probe this path as follows: **(1)** As a feature curve is a point set whose deformation has a high degree of freedom, directly optimizing the per-point offsets often leads to undesired self-intersection and spike artifacts. To better regularize the deformation of curves, we introduce an *intersection-free curve deformation* method to maintain the order of feature curves. **(2)** We optimize the 3D feature curves using 2D projection loss measured by the estimated 2D visible curves, where the key challenge is to accurately compute the visibility of curves. To address this problem, we propose a *surface-aware curve visibility estimation* method based on the implicit garment surface and z-buffer. **(3)** To ensure the accuracy of curve visibility estimation during the optimization process, the curves should always be right on the garment surface. We therefore introduce a *progressive curve and surface evolution* strategy to jointly update the curves and surface while imposing the on-surface regularization for curves.

To summarize, the main contributions of this work are:

- We introduce **REC-MV**, to our best knowledge, the **first** method to reconstruct dynamic and open loose garments from the monocular video.
- We propose a new approach for joint optimization of explicit feature curves and implicit garment surface from monocular video, based on carefully designed intersection-free curve deformation, surface-aware curve visibility estimation, and progressive curve and surface evolution methods.
- Extensive evaluations on casually captured monocular videos demonstrate that our method outperforms existing methods.

* feature curves of the garment (*e.g.*, necklines, hemlines) can provide critical cues for determining the shape contours of the garment.

2. Related Work

Human Reconstruction from Single-view Image. Traditional methods for human reconstruction often adopt a parametric human model (*e.g.*, SMPL [32] or SCAPE [4]) and can only recover a naked 3D body [23, 24]. To increase the surface details, free-form deformations can be applied to the mesh vertices to model small geometry variations caused by the clothing [2, 3, 26, 43, 52].

Recent methods propose to utilize implicit surface representations [34, 38] to reconstruct 3D clothed human with an arbitrary topology. Specifically, PIFu and PIFuhd [44, 45] extract pixel-aligned spatial features from images as the input for implicit surface function for occupancy prediction. Follow-up methods then integrates 3D-aligned features to improve the results [6, 15–18, 56, 58]. As these methods only consider single-image reconstruction, they cannot produce temporally consistent results for video input.

Human Reconstruction from Monocular Video. Inspired by the success of neural rendering methods [35, 37, 57] in scene reconstruction, many methods have been proposed to reconstruct 3D human from sparse-view [31, 41, 53, 55, 60] or monocular [19, 47, 49] videos.

Anim-NeRF [8], Neuman [21] and HumanNeRF [49] introduce methods based on neural radiance field (NeRF) [35] to reconstruct an animatable avatar from monocular video. These methods transform a 3D point in the observation space to the canonical space by inverse-skinning, and then perform volume rendering in the canonical space. A-NeRF [47] additionally adopt a skeleton-relative encoding strategy. AvatarCap [29] proposes a monocular human volumetric capture method, but requires reconstructing an avatar from multiple 3D scans in advance.

Garment Reconstruction from Images. Reconstructing garment mesh from images enables many applications like virtual try-on and content creation. Existing methods reconstruct the clothing as a separate layer on top of the body [7, 22, 25, 42, 48, 51]. Among them, several methods address the challenging problem of garment reconstruction from single-view image [5, 10, 20, 36, 61, 62]. MGN [5] learns a per-category parametric model from a large-scale clothing dataset. BCNet [20] first reconstructs a coarse template and then refines the surface details with a displacement network. AnchorUDF [59] adopts the unsigned distance field (UDF) [9] to represent the open surface mesh. SMPlicit [10] proposes a generative model to reconstruct layered garments from a single image. DeepFasion3D [61] reconstructs the surface with occupancy network [34] and applys non-rigid ICP to register the clothing template. ReEF [62] registers explicit clothing template to the implicit field learned from pixel-aligned implicit function. However, as these single-image methods do not consider clothing motion, they are not suitable for dynamic gar-

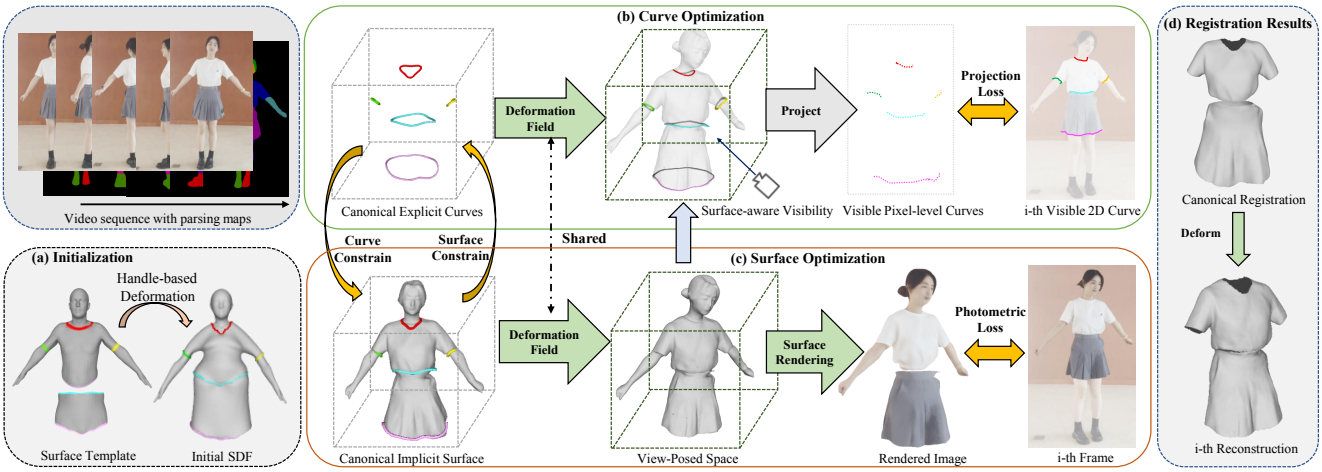


Figure 2. **Overview of the proposed REC-MV**. (a) Starting from a surface template, we initialize the canonical curves by solving Eq. (3), and apply a handle-based deformation to initialize the canonical implicit surface. (b) Given an i-th frame, canonical curves are deformed to the camera view space to compute the projection loss based on the surface-aware visibility estimation. (c) Similarly, the canonical surface is deform to the camera view to compute the photometric loss by differentiable rendering. The curves and surface are jointly optimized to enable a progressive co-evolution. (d) Last, the open garment meshes can be extracted by template registration in the canonical space.

ment reconstruction.

Among methods related to garment reconstruction from videos, Li *et al*. [28] introduce a method to learn physics-aware clothing deformation from monocular videos, but assumes the template scans for the body and clothing are provided [13]. Garment Avatar [14] proposes a multi-view patterned cloth tracking algorithm, requiring the subject to wear clothing with specific patterns. SCARF represents the layered clothing using radiance field [11] on top of the SMPL-X model [40] from monocular video. In contrast, our method first reconstructs the explicit 3D garment curves and surfaces, and then extracts the garment mesh via template registration.

3. Method

Given a monocular video with N_i frames depicting a moving person $\{I_t|t = 1, \ldots, N_i\}$, REC-MV aims to reconstruct high-fidelity and space-time coherent *open* garment meshes. This is a challenging problem as it requires a method to simultaneously capture the shape contours, local surface details, and the motion of the garment.

Observing that feature curves (*e.g*., necklines, hemlines) provide critical cues for determining the shape contours of garment [61] and implicit signed distance function (SDF) can well represent a detailed *closed* surface [19], we propose to first optimize the explicit 3D feature curves and implicit garment surfaces from the video, and then apply non-rigid clothing template registration to extract the open garment meshes (see Fig. 2).

Preprocessing. We generate the initial shape parameter $\boldsymbol{\beta}$, camera intrinsic π, and per-frame SMPL [32] pose parameters $\{\boldsymbol{\theta}_t|t = 1, \ldots, N_i\}$ using Videoavatar [3]. To identify the garment regions in 2D images, we apply the existing garment parsing method [27] to estimate the garment masks. Our method also requires 2D visible curves $\zeta = \{\zeta_{l,t}|l = 1, \ldots N_l, t = 1, \ldots, N_i\}$ for 3D curve recovery, where N_l denotes the number of curves. Note that the 2D visible curves can be automatically produced by parsing boundaries of the garment mask (more details in the supplementary material).

Overview. To utilize the information that exists in the entire video for dynamic garment reconstruction, we represent the explicit feature curve and implicit garment surface in the canonical space (Sec. 3.1). For a specific time step, we adopt the skeleton-based skinning and non-rigid deformation modeling to map the canonical curves and surfaces to the camera view space (Sec. 3.2). As the given 2D curves only contain visible points, to optimize the 3D feature curves from 2D projection error, we propose a surface-aware approach to compute the visibility of the 3D feature curve based on z-buffer (Sec. 3.3). In terms of implicit surface optimization, we minimize the photometric loss between the rendered and input image based on the differentiable surface rendering technique (Sec. 3.4). Then the adopted loss functions for joint optimization of curves and surfaces are described (Sec. 3.5). Last, the open garment meshes can be extracted by registering an explicit garment template to the recovered curves and implicit surfaces in the canonical space (more details in the supplementary material). Then the garment meshes can be deformed based on the SMPL poses.

3.1. Feature Curve and Surface Representation

Explicit Surface Template. Following DeepFashion3D [61], we employ several surface templates, each contains a pre-defined set of 3D feature curves $\mathbf{L} = \{\mathbf{L}_i | i = 1, \ldots, N_l\}$ extracted from the garment boundaries, where N_l is the number of feature curves (see our supplementary materials for more details). The surface templates will be used for garment surface initialization and the pre-defined feature curves will be used for curve initialization[†].

Intersection-free Curve Deformation. A straightforward idea is to represent a feature curve as a discrete point set, and directly estimate the 3D deformation offset for each point during optimization. However, this unstructured curve representation struggles to maintain the order of the points and often generate spike artifacts due to the high degree of freedom of the deformation.

To address this issue, we introduce a novel intersection-free curve deformation method, in which the point's deformation at each step is controlled by the curve center and two orthogonal directions (see Fig. 3 for illustration). Formally, given a curve $\mathcal{C}$ of N_p points with center $\mathbf{p_c}$, the updated position of i-th point $\mathcal{C}(i)$ is defined as

$$\mathcal{C}'(i) = \mathbf{p_c} + S_i^d \mathbf{n_i^d} + S_i^c \mathbf{n^c}, \tag{1}$$

where $\mathbf{n_i^d}$ is the direction from the curve center to the current point $\mathcal{C}(i)$, and $\mathbf{n^c} = \frac{1}{N_p - 1} \sum_{i=1}^{N_p} (\mathbf{n_i^d} \times \mathbf{n_{i-1}^d})$ is the direction perpendicular to the current feature curve plane. $S_i^d \in \mathbb{R}$ and $S_i^c \in \mathbb{R}$ are learnable parameters specifying the step size of the deformation.

The proposed intersection-free curve deformation can well preserve the order of points in the curve, which largely reduced the difficulty of optimization compared to the direct offset estimation approach.

Implicit SDF in Canonical Space. Unsigned distance field (UDF) [9] is an implicit function that can represent an open surface. However, as UDF is not differentiable at points close to the surface, it is non-trivial to integrate UDF with differentiable surface rendering to take advantage of supervision from 2D photometric loss. We therefore adopt the SDF to represent a closed garment surface for surface geometry recovery, followed by garment template registration to extract the open surface.

It is common to represent the whole surface with a single SDF for human reconstruction [19]. However, as our goal is to reconstruct separate clothes, using a single SDF to represent both the upper clothes and bottom clothes (*e.g.*, skirt) increases the difficulty of template registration (*i.e.*, splitting the upper and bottom clothes requires highly accurate waist curves).

[†] including templates for uppers, dresses, coats, pants, and skirts.

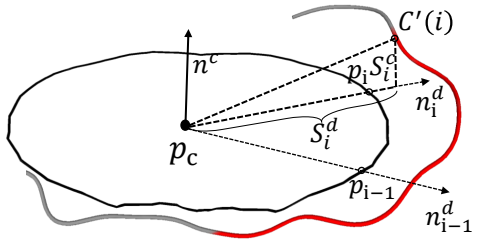


Figure 3. Illustration of the intersection-free curve deformation.

To enable better template registration, we consider three different surface types (*i.e.*, *upper-clothing*, *bottom-clothing*, and *upper-bottom*) according to the garment types, and represent each surface type as the zero-isosurface of an independent SDF in the canonical space. The SDF is expressed by an MLP f with learnable weights η:

$$S(\eta) = \{\mathbf{p} \in \mathbb{R}^3 | f(\mathbf{p}; \eta) = 0\}.$$

For the sake of simplicity and without loss of generality, we illustrate our method in reconstructing a single surface type later in this section.

3.2. Skinning Based Motion Modeling

We model large body motions by linear blend skinning (LBS) transformation based on the SMPL [32] model, and utilize a non-rigid deformation field to account for fine-grained deformations.

Skinning Transformation. Given a SMPL body with shape parameter $\boldsymbol{\beta}$ and a pose parameter $\boldsymbol{\theta}_i$ in i-th frame, a point $\mathbf{p}$ on the body surface in canonical space with skinning weights $w(\mathbf{p})$ can be warped to camera view space via skinning transformation $\mathcal{W}$.

Notably, the skinning weights $w(\mathbf{p})$ are only defined for points on the SMPL surface. To warp arbitrary points in the canonical space to camera view, we use the diffused skinning strategy [30] to propagate the skinning weights of SMPL body vertices to the entire canonical space, and store the weights in a voxel grid of size $256 \times 256 \times 256$. Then we can obtain the skinning weights by trilinear interpolation.

Non-rigid Deformation. Skinning deformation enables the garment surface to deform in a way consistent with the body's large-scale motion [16]. However, the motion of details and garment parts that are far away from body cannot be fully represented by skinning transformation [19]. Hence, a non-rigid deformation MLP is used to model these fine-grained changes. Specifically, we design an MLP $\mathcal{D}$ with learnable parameters ϕ to model garment surface's non-rigid deformation:

$$\mathbf{p}' = \mathcal{D}(\mathbf{p}, \mathbf{h}, E(\mathbf{p}); \phi), \tag{2}$$

where $\mathbf{p}'$ is the deformed point of the input point $\mathbf{p}$ in the canonical space, $\mathbf{h}$ is the latent code of the current frame,

and $E(\mathbf{p})$ of $\mathbf{p}$ is the position encoding [35] to represent the high-frequency information of spatial points.

Finally, combining $\mathcal{D}$ with skinning transformation field $\mathcal{W}$, we could define a deformation field $\Phi(\cdot) = \mathcal{W}(\mathcal{D}(\cdot))$ to warp any points in the canonical space to the camera view.

3.3. 3D Feature Curves from 2D Projections

The 3D feature curve will be optimized by minimizing the distance between its 2D projection on the image plane and the provided 2D visible curves. The key challenge here is how to compute the visibility of the 3D curves in the camera view. We first introduce a curve initialization strategy based on rigid transformation, and then propose a surface-aware curve visibility estimation method to support accurate non-rigid curves optimization.

Feature Curve Initialization. We start from the predefined feature curve sets $\mathbf{L} = \{\mathbf{L}_i | i = 1, \ldots, N_l\}$ provided in the garment template. To reduce the difficulty of curve optimization, we perform a rigid curve initialization by directly minimizing the Chamfer Distance (CD) between the projected curves on the camera view space and the corresponding visible 2D curves ζ as

$$s, \mathbf{t}, \mathbf{R} = \underset{s,\mathbf{t},\mathbf{R}}{\arg\min}\, \mathrm{CD}\left(\Pi\left(\mathcal{W}(\bar{\mathbf{L}}_i)\right), \zeta_i\right), \tag{3}$$

$$\bar{\mathbf{L}}_i = s\mathbf{R}(\mathbf{L}_i) + \mathbf{t}, \tag{4}$$

where Π is the projection matrix, $\bar{\mathbf{L}}_i$ is the transformed feature curve. $\mathbf{t} \in \mathbb{R}^3$, $\mathbf{R} \in SO(3)$, and $s \in \mathbb{R}$ are the optimized translation, rotation, and scaling parameters, respectively.

In our implementation, we execute 150 gradient descent iterations to solve the rigid transformation parameters. After rigid optimization, we set $\bar{\mathbf{L}}$ as the initial position for the feature curve sets $\{\mathcal{C}_i | i = 1, \ldots, N_l\}$ for later non-rigid optimization.

Surface-aware Curve Visibility Estimation. As the 2D feature curve ζ only contains visible points, it is essential to identify the visible points of the 3D curve $\mathcal{C}$ in camera space. A naive solution is to consider a point $\mathcal{C}(i)$ as visible if the cosine similarity between the view direction $\mathbf{v}$ and $\mathbf{n}_\mathbf{i}^\mathbf{d}$ (*i.e.*, the direction from curve center to the i-th point) in view-pose is less than 0. However, this approach will produce wrong judgments when a curve is occluded by other body parts.

To tackle this problem, a surface-aware curve visibility estimation method is proposed. Specifically, we generate an explicit mesh $\mathbf{T_s}$ from implicit surface $S(\eta)$ in canonical space via marching cube [33]. Next, we deform $\mathbf{T_s}$ to camera view space via the deformation field $\Phi(\mathbf{T_s})$. Then, we can check if a feature curve point $\Phi(\mathcal{C}(i))$ is occluded by the explicit mesh in view space based on z-buffer:

$$V_{\mathcal{C}(i)} = \mathrm{zbuffer\_test}(\Phi(\mathcal{C}(i)), \Phi(\mathbf{T_s})). \tag{5}$$

However, we find that the 3D curve $\mathcal{C}$ might sometimes move outside or have a scale larger than the explicit mesh $\mathbf{T_s}$, there will be some errors if only depending on the z-buffer testing between $\mathcal{C}(i)$ and $\mathbf{T_s}$. We therefore make use of the SMPL surface to improve the visibility estimation, by checking if the nearest point of $\mathcal{C}(i)$ on the SMPL body is occluded in the camera view space in a similar way. Note that this is feasible as in our intersection-free curve deformation, the correspondences between $\mathcal{C}(i)$ and its nearest vertice in the SMPL body are almost unchanged during optimization. Then a curve point is considered as visible if it passes both visibility checks.

3.4. Progressive Curve and Surface Co-evolution.

The surface of the garment is represented by the implicit SDF. As the feature curve visibility estimation depends on the garment surface, the curves and surface have to evolve consistently. To ensure the accuracy of curve visibility during the optimization process, we jointly optimize the curves and surface while imposing a regularization that the curves lie on the zero-isosurface of the SDF. The implicit surface is minimized by the photometric loss based on differentiable surface rendering.

Curve-aware Surface Initialization. A good initialization for the implicit SDF $S(\eta)$ can reduce the optimization difficulty and improve the performance, especially for the long skirt and dress. Thanks to our curve-aware garment representation, we can utilize the initialized feature curve $\bar{\mathbf{L}}$ computed in Eq. (4) to enable a better shape initialization. Specifically, we apply a handle-based deformation [46] to deform a surface template such that its feature curves are aligned with $\bar{\mathbf{L}}$. Then, we apply IGR [12] to initialize the implicit surface $S(\eta)$ by fitting the deformed template.

Differentiable Surface Rendering. To reconstruct high-fidelity geometry, following the SelfRecon [19], we find the intersection points $\mathbf{p}$ on the surface and make them differentiable (more details can be found in supplementary).

After obtaining the intersection points $\mathbf{p}$, we compute its gradient $\mathbf{n_p} = \nabla f(\mathbf{p}; \eta)$ and transform the camera view to canonical space as $\mathbf{v_p}$ by the Jacobian matrix of the deformed point $\Phi(\mathbf{p})$ (more details can be found in supplementary). To better account for the changes in the illumination, we also take a per-frame latent code $\mathbf{z}$ as input to the color rendering network f_c. Then, the surface color $C_\mathbf{p}$ of point $\mathbf{p}$ can be computed as

$$C_\mathbf{p} = f_c(\mathbf{p}, \mathbf{n_p}, \mathbf{v_p}, \mathbf{z}, E(\mathbf{p}); \psi). \tag{6}$$

3.5. Loss Function

The overall loss function consists of two parts, one part is for the feature curves optimization and the other is for garment surfaces optimization.

3.5.1 Explicit Feature Curve Loss

The optimization of feature curves relies on the 2D projection loss, a curve slope regularization loss, and an on-surface regularization loss that ensures the feature curves are on the garment surface.

Feature Curve Projection Loss. Given SMPL pose parameter $\boldsymbol{\theta}_i$ and the camera projection matrix Π, we warp predicted feature curve $\mathcal{C}$ to camera view space via deformation field Φ, and compute project loss $\mathcal{L}_{proj}$ measured by 2D visible curves ζ using Chamfer Distance (CD):

$$\mathcal{L}_{proj} = \mathrm{CD}(V_{\mathcal{C}} \otimes \Pi(\Phi(\mathcal{C})), \zeta) \tag{7}$$

where $V_{\mathcal{C}}$ is the visibility mask of curves, and symbol $\otimes$ indicates the mask selection operator.

Feature Curve Slope Regularization. To maintain the curvature of 3D curve $\mathcal{C}$, we design a slope loss $\mathcal{L}_{slop}$ to regularize that the slope is consistent between adjacent points

$$\mathcal{L}_{slop} = \sum_{i=1}^{N_p} (1 - cos < \mathbf{s}_{i+1}, \mathbf{s}_i >) \tag{8}$$

where $\mathbf{s}_i = \mathcal{C}(i+1) - \mathcal{C}(i)$, N_p is the point number in the curve, and $cos <>$ is the cosine similarity function.

On-surface Regularization. In addition, the feature curves are required to be on the corresponding garment surface. Hence, we introduce an as near as possible loss $\mathcal{L}_{anap}$ as:

$$\mathcal{L}_{anap} = \sum_{i=1}^{N_p} |f(\mathcal{C}(i); \eta)| \tag{9}$$

The overall explicit feature curve loss can be written as:

$$\mathcal{L}_{curve} = \lambda_{proj}\mathcal{L}_{proj} + \lambda_{slop}\mathcal{L}_{slop} + \lambda_{anap}\mathcal{L}_{anap} \tag{10}$$

where $\lambda_{proj}, \lambda_{slop}$ and λ_{anap} are loss weights.

3.5.2 Garment Surface Loss

For a monocular video with N_i frames, the learnable parameter in implicit surface reconstruction is denoted as $\boldsymbol{\Theta}$:

$$\boldsymbol{\Theta} = \{\eta, \phi, \psi\} \cup \{\mathbf{h}_i, \mathbf{z}_i | i = 1, \ldots, N_i\} \tag{11}$$

Surface Rendering Loss. For a pixel within the garment mask, we compute the ray's intersection point $\mathbf{p}$ on the canonical surface $S(\eta)$ and apply surface rendering network to predict the color $C_{\mathbf{p}}$ (see Eq. (6)). Then the photometric loss can be computed as

$$\mathcal{L}_{RGB} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{p} \in \mathcal{R}} |C_{\mathbf{p}}(\boldsymbol{\Theta}) - I_{\mathbf{p}}|, \tag{12}$$

where $\mathcal{R}$ is the sample point set, $I_{\mathbf{p}}$ is the corresponding ground-truth pixel color from the input images.

Mask-guided Implicit Consistency Loss. To better optimize implicit surface, following SelfRecon [19], we periodically extract explicit surface meshes $\mathbf{T_s}$ in canonical space from SDF f and use a differentiable renderer [50] to iteratively optimize $\mathbf{T_s}$ by a mask loss using the surface mask. Then the updated explicit surface $\hat{\mathbf{T}}_\mathbf{s}$ will be used to supervise the implicit SDF f as

$$\mathcal{L}_{mcons} = \frac{1}{|\hat{\mathbf{T}}_\mathbf{s}|} \sum_{\mathbf{p} \in \hat{\mathbf{T}}_\mathbf{s}} |f(\mathbf{p}; \eta)|. \tag{13}$$

Curve-guided Implicit Consistency Loss. We find that the explicit mesh $\hat{\mathbf{T}}_\mathbf{s}$ updated by the mask loss might contain holes or even collapse in some surface areas, which will harm the learning of implicit surface (see in Fig. 9). To address this issue, we design an explicit curve and surface consistency loss. Specifically, for a specific feature curve $\mathcal{C}$ that belongs to two implicit surfaces (*e.g.*, waist curve belongs to both the upper-clothing and bottom-clothing), we generate its closed surface $\mathbf{T}_\mathcal{C}$ and then sample N_a points from $\mathbf{T}_\mathcal{C}$ to constrain the implicit SDF f as

$$\mathcal{L}_{ccons} = \frac{1}{|\mathbf{T}_\mathcal{C}|} \sum_{\mathbf{p} \in \mathbf{T}_\mathcal{C}} |f(\mathbf{p}; \eta)|. \tag{14}$$

Common Implicit Loss. Eikonal loss $\mathcal{L}_{eik}$ [12] is included to make the implicit function the signed distance function. To avoid distortion of non-rigid transformation, a rigid loss [39] $\mathcal{L}_{arap}$ is computed to constrain the non-rigid deformation. We also compute normal loss $\mathcal{L}_{norm}$ in canonical space to further refine the surface [19]. Moreover, we compute the skeleton smoothness loss [54] to reduce the high-frequency jitter of SMPL poses among frames (more details can be found in supplementary).

The overall implicit surface loss can be written as:

$$\begin{aligned} \mathcal{L}_{ims} = \mathcal{L}_{RGB} + \lambda_{mcons}\mathcal{L}_{mcons} + \lambda_{ccons}\mathcal{L}_{ccons} \\ \lambda_{arap}\mathcal{L}_{arap} + \lambda_{eik}\mathcal{L}_{eik} + \lambda_{norm}\mathcal{L}_{norm}, \end{aligned} \tag{15}$$

where λ_{arap}, λ_{mcons}, λ_{ccons}, λ_{eik}, and λ_{norm} are the loss weights.

4. Experiments

Since there is no existing method for open garment meshes reconstruction from monocular videos, we compare with three state-of-the-art single-image methods, namely BCNet [20], ClothWild [36], and ReEF [62].

4.1. Evaluation on Synthetic Dataset

Since there is no public real dataset for evaluating dynamic garment reconstruction, we adopt four video sequences from the synthetic data generated by SelfRecon [19] for quantitative evaluation.

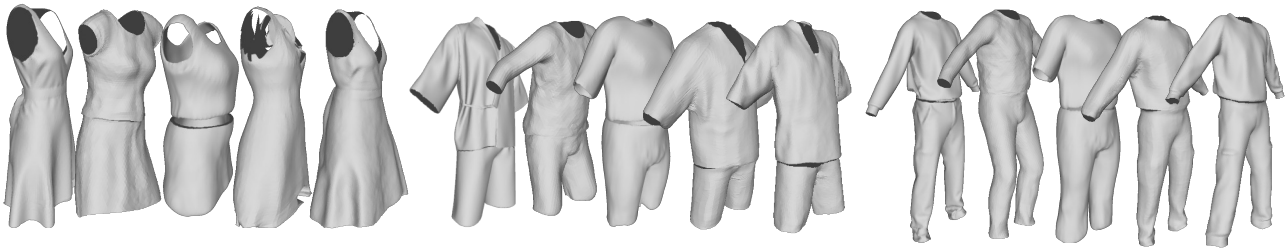


Figure 4. Qualitative comparison on the synthetic dataset. From left to right in each example: the ground-truth mesh, results of BCNet [20], ClothedWild [36], ReEF [62], and Ours.

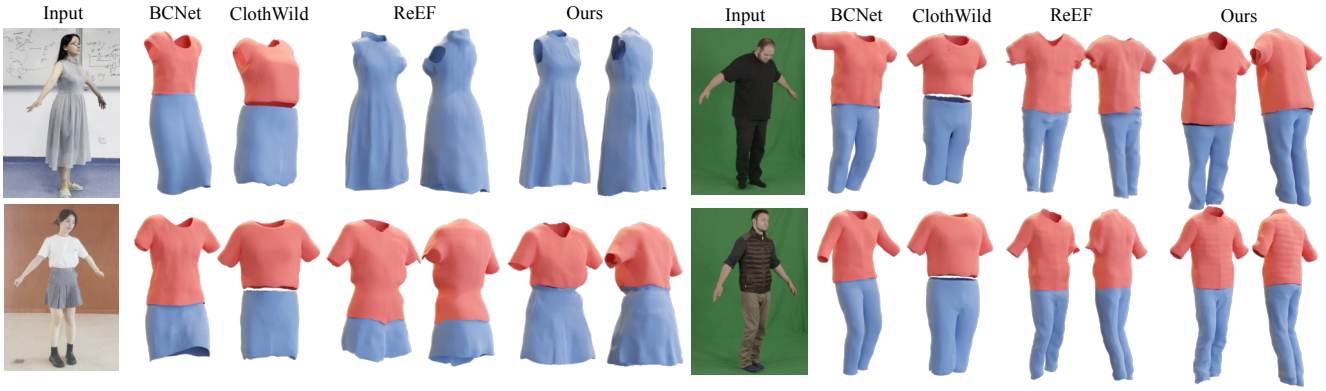


Figure 5. Qualitative comparison on real datasets between BCNet [20], ClothWild [36], ReEF [62], and our method. Upper clothes are visualized in red color, while bottom clothes and dresses are visualized in blue color. Note that BCNet and ClothWild cannot model dresses.

We first employ Blender [1] to extract the ground-truth garment mesh from the provided clothed human mesh of the first frame. To measure the *accuracy* of the reconstructed meshes, we compute the Chamfer distance (CD) between the ground-truth and estimated meshes. To evaluate the *temporal consistency* of the reconstructed meshes for the video sequence, we measure the consistency of corresponding vertices (CCV), which is the root mean square error of the corresponding vertices distances in adjacent frames.

We test our method and the baseline methods on these four video sequences. Table 1 shows that our method achieves the best results in the metrics of CD and CCV on all four videos, demonstrating the effectiveness of our method in reconstructing accurate and temporally consistent dynamic garment meshes. From the results of high errors in the CCV, we can clearly see that single-image methods fail to maintain the consistency of the reconstruction for the video input. Figure 4 compares the visual results, in which our method produces detailed and accurate garments that are mostly close to the ground-truth surfaces.

4.2. Evaluation on Real-world Videos

We then qualitatively evaluate our method on the PeopleSnapshot [3] and a dataset captured by ourselves. These testing videos include a diverse variety of garments categories, including upper-cloth, dress, coats, pants, and skirts.

Table 1. Quantitative results on four synthetic sequences. We compare the Chamfer distance (CD) between the ground-truth and reconstructed surfaces (in cm), as well as the consistency of corresponding vertices (CCV) between adjacent frames.

Method	Female1		Female3		Male1		Male2	
	CD	CCV	CD	CCV	CD	CCV	CD	CCV
BCNet [20]	3.184	7.201	3.447	6.186	2.929	8.604	5.234	7.128
ClothedWild [36]	2.424	-	2.075	-	2.782	-	3.980	-
ReEF [62]	1.810	3.782	1.924	4.322	2.005	6.794	2.865	3.579
Ours	**1.804**	**0.597**	**1.641**	**1.064**	**1.736**	**0.484**	**1.812**	**0.433**

Figure 5 shows the visual comparisons. The results of the baseline methods are predicted using a single image as input. Our method can faithfully reconstruct the layouts and surface details of the garments. In contrast, BCNet [20] and ClothWild [36] cannot accurately predict the garment layouts and produce over-smooth surfaces.

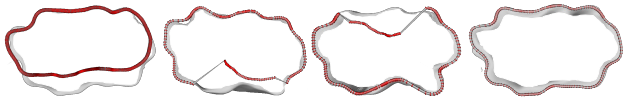
We also demonstrate our dynamic reconstruction results in Fig. 6. We can see that our method can produce space-time coherent results for different garment types (including the challenging dresses) from monocular videos, which is difficult to achieve with single-image methods.

4.3. Ablation Study

We next conduct ablation study for different components of our method (more results in our supplementary material).

Curve Visibility Estimation. As shown in Fig. 7, sim-

Figure 6. Dynamic garment reconstruction results of our method. Each row shows the reconstruction of four frames in a monocular video.



(a) Curve Normal (b) w/o Implicit Surface (c) w/o SMPL (d) Surface-aware

Figure 7. Ablation study of curve visibility estimation method.

ply using normal direction for visibility estimation leads to worse results, while using both the implicit SDF and SMPL surfaces for z-buffer testing produces the best result.

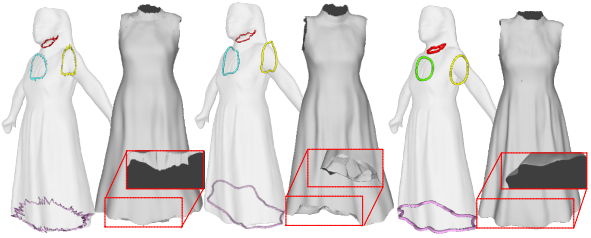
Explicit Curve Losses. Figure 8 (a) shows that without the curve slop loss $\mathcal{L}_{slop}$, the optimized curves will contain noise and artifacts. As shown in Fig. 8 (b), the proposed on-surface regularization (*i.e.*, $\mathcal{L}_{anap}$) can well constrain the curves to be on the surface and produce much more accurate fitting results, demonstrating the implicit surface helps the optimization of curves.

Curve-guided Consistency Loss. To improve the optimization of the implicit surface, we use curves to regularize the surface. Figure 9 shows that this regularization effectively improves the surface geometry, verifying that curves benefit the optimization of surfaces.

5. Conclusion

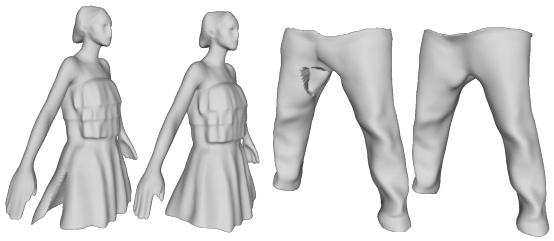
We have presented a new framework for dynamic garment reconstruction from monocular videos, by formulating this task as an optimization problem of dynamic 3D curves and surface recovery, followed by garment template registration. To solve this problem, we introduce a novel approach, called REC-MV, to jointly optimize the curves and surface from 2D supervision in a progressive co-evolution manner. Experimental results show that our method can reconstruct high-fidelity dynamic garments meshes with open boundaries, significantly outperforming existing methods.

Limitations. Our method can only reconstruct common garment categories whose contours can be represented by



(a) w/o $\mathcal{L}_{slop}$ (b) w/o $\mathcal{L}_{anap}$ (c) Both

Figure 8. Ablation study of the explicit curve losses.



w/o $\mathcal{L}_{ccons}$ w/ $\mathcal{L}_{ccons}$ w/o $\mathcal{L}_{ccons}$ w/ $\mathcal{L}_{ccons}$

Figure 9. Results of w/o and w/ the curve-guided consistency loss.

feature curves. Additionally, our method requires the moving person to be observed from different angles.

Acknowledgement. The work was supported in part by NSFC with Grant No. 62293482, the Basic Research Project No.HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone. It was also partially supported by Shenzhen General Project with No.JCYJ20220530143604010, the National Key R&D Program of China with grant No.2018YFB1800800, by NSFC No. 62202409, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No.2017ZT07X152 and No.2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No.2022B1212010001), and by Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No.ZDSYS201707251409055). It was also sponsored by CCF-Tencent Open Research Fund.

References

[1] Blender. https://www.blender.org/. 2021. 7

[2] 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 *CVPR*, 2019. 2

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 2, 3, 7

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *TOG*, 2005. 2

[5] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *ICCV*, 2019. 1, 2

[6] 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 *CVPR*, 2022. 2

[7] Andrés Casado-Elvira, Marc Comino Trinidad, and Dan Casas. Pergamo: Personalized 3d garments from monocular video. *arXiv e-prints*, 2022. 2

[8] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 1, 2

[9] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *NeurIPS*, 2020. 2, 4

[10] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 1, 2

[11] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. *arXiv preprint arXiv:2210.01868*, 2022. 3

[12] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Machine Learning and Systems*, 2020. 5, 6

[13] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 3

[14] Oshri Halimi, Fabian Prada, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, and Yaser Sheikh. Garment avatars: Realistic cloth driving using pattern registration. *arXiv preprint arXiv:2206.03373*, 2022. 3

[15] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *NeurIPS*, 2020. 2

[16] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *ICCV*, 2021. 2, 4

[17] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *CVPR*, 2021. 2

[18] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 2

[19] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6

[20] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *ECCV*, 2020. 1, 2, 6, 7

[21] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, 2022. 2

[22] N. Jin, Y. Zhu, Z. Geng, and R. Fedkiw. A pixel-based framework for data-driven clothing. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2020. 2

[23] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *ICCV*, 2018. 2

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[25] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *ECCV*, 2018. 2

[26] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *International Conference on 3D Vision (3DV)*, 2019. 2

[27] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *TPAMI*, 2020. 3

[28] Yue Li, Marc Habermann, Bernhard Thomaszewski, Stelian Coros, Thabo Beeler, and Christian Theobalt. Deep physics-aware inference of cloth deformation for monocular human performance capture. In *3DV*, 2021. 3

[29] Zhe Li, Zerong Zheng, Hongwen Zhang, Chaonan Ji, and Yebin Liu. Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In *ECCV*, 2022. 1, 2

[30] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *ECCV*, 2022. 4

[31] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *TOG*, 2021. 2

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 2015. 2, 3, 4

[33] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH*, 1987. 5

[34] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks:

Learning 3d reconstruction in function space. In *ICCV*, 2019. 2

[35] 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. In *ECCV*, 2020. 1, 2, 5

[36] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction in the wild. In *ECCV*, 2022. 2, 6, 7

[37] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 1, 2

[38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *ICCV*, 2019. 2

[39] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 6

[40] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 3

[41] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2

[42] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *TOG*, 2017. 2

[43] Ma Qianli, Yang Jinlong, Ranjan Anurag, Pujades Sergi, Pons-Moll Gerard, Tang Siyu, and J.Black Michael. Learning to dress 3d people in generative clothing. In *CVPR*, 2020. 2

[44] 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 *ICCV*, 2019. 2

[45] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2

[46] Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. Laplacian surface editing. In *Eurographics/ACM SIGGRAPH symposium on Geometry processing*, 2004. 5

[47] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021. 1, 2

[48] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *ECCV*, 2020. 2

[49] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 1, 2

[50] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, 2020. 6

[51] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *arXiv preprint arXiv:2206.15470*, 2022. 2

[52] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica K. Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular RGB video. In *3DV*, 2020. 2

[53] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *NeurIPS*, 2021. 2

[54] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *TOG*, 2018. 6

[55] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 2

[56] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *CVPR*, 2021. 2

[57] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, 2020. 1, 2

[58] Zheng Zerong, Yu Tao, Liu Yebin, and Dai Qionghai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *TPAMI*, 2021. 2

[59] Fang Zhao, Wenhao Wang, Shengcai Liao, and Ling Shao. Learning anchored unsigned distance functions with gradient direction alignment for single-view garment reconstruction. In *ICCV*, 2021. 2

[60] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *CVPR*, 2021. 2

[61] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *ECCV*, 2020. 1, 2, 3, 4

[62] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *CVPR*, 2022. 1, 2, 6, 7