

D³-Human: Dynamic Disentangled Digital Human from Monocular Video

Honghu Chen Bo Peng Yunfan Tao Juyong Zhang*

University of Science and Technology of China

{honghuc@mail., ustcpbb@mail., taoyf@mail., juyong@}ustc.edu.cn

Abstract

We introduce *D³-Human*, a method for reconstructing *Dynamic Disentangled Digital Human* geometry from monocular videos. Past monocular video human reconstruction primarily focuses on reconstructing undecoupled clothed human bodies or only reconstructing clothing, making it difficult to apply directly in applications such as animation production. The challenge in reconstructing decoupled clothing and body lies in the occlusion caused by clothing over the body. To this end, the details of the visible area and the plausibility of the invisible area must be ensured during the reconstruction process. Our proposed method combines explicit and implicit representations to model the decoupled clothed human body, leveraging the robustness of explicit representations and the flexibility of implicit representations. Specifically, we reconstruct the visible region as SDF and propose a novel human manifold signed distance field (hmSDF) to segment the visible clothing and visible body, and then merge the visible and invisible body. Extensive experimental results demonstrate that, compared with existing reconstruction schemes, *D³-Human* can achieve high-quality decoupled reconstruction of the human body wearing different clothing, and can be directly applied to clothing transfer and animation production. Code is available at <https://ustc3dv.github.io/D3Human/>.

1. Introduction

Clothed human body reconstruction has long been a research focus in the fields of graphics and computer vision, and has a wide range of applications in many fields such as virtual reality, augmented reality, holographic communication, film production, and game development. Compared to film-level reconstructions requiring numerous cameras and artists for modeling, reconstructing high-quality clothed human bodies from monocular videos holds more practical value for general users. In scenarios such as telepresence and virtual try-ons, the 3D avatars used should be

easily accessible, visually realistic, and easy to edit, including modifications to clothing and posture. Therefore, how to reconstruct high-fidelity, decoupled representations of clothed human bodies using monocular video remains a long-standing research problem. Through decoupled reconstruction, clothing can be separated from the human body, enabling efficient adjustment and editing of different clothing styles, postures, and body shapes. This decoupling not only enhances the flexibility and practicality of 3D reconstruction, but also improves the realism of detailed features, providing greater potential for personalization and dynamic interaction of virtual characters.

We aim to develop a method to decouple and reconstruct the clothed human body from monocular videos. However, this is a very challenging task because 1) monocular videos only provide single-view 2D image information and lack direct 3D depth perception, 2) real captured videos contain various clothing styles, irregular textures, and complex human poses, 3) body parts occluded by clothes are not visible in the input video, which also poses a great challenge to reconstruction. Existing methods can be divided into explicit expression methods and implicit expression methods. Among them, explicit expression methods usually rely on pre-acquired templates. Some methods [16, 17, 54] use scanners, while other methods [2, 28, 39] rely on parametric models [20, 32], and the reconstruction quality mainly depends on the representation ability of the model. Implicit representation methods [9, 14, 21, 23, 40] use NeRF [35] or SDF [38] to model the clothed human body, but they usually produce an inseparable whole body or exhibit average geometric quality.

We propose a decoupled human body reconstruction scheme, named *D³-Human* (Dynamically Disentangled Digital Human), which combines explicit and implicit representations to address the main challenges of template generation and the dynamic deformations. In decoupled human body reconstruction, generating clothing templates is particularly challenging. Traditional methods, such as those based on parametric models [28] or feature lines [43], rely heavily on priors, limiting the types of clothing they can represent. Although implicit unsigned distance field (UDF)

*Corresponding author

representations [5, 13, 29, 31] provide some solutions, they perform poorly when single-view supervision is limited (see experiments 4.3). For the visible region, inspired by G-Shell [30] and DMTet [50], we define an optimizable human manifold signed distance field (hmSDF) on the non-decoupled clothed human surface to separate clothing from the body. For the invisible regions of the body, we adopt the explicit representation of the corresponding regions of the SMPL [32] model to ensure the plausibility of the body shape and seamless integration with the visible region. This approach enables detailed and decoupled reconstruction of clothed human bodies.

We reconstruct different human bodies wearing various outfits based on monocular videos, to demonstrate the capabilities of our method. Compared to existing methods that require over a day [21, 43, 52] for reconstruction, D³-Human reconstructs a decoupled template of clothing and body in a much shorter time (about 20 minutes) and completes the full sequence within several hours, achieving competitive reconstruction accuracy. Furthermore, we demonstrate examples of applications in animation production and clothing transfer to showcase the wide-ranging applicability of the decoupled representation. In summary, the contributions of this paper include the following aspects:

- A hybrid reconstruction method combining explicit and implicit representations is proposed, which is capable of reconstructing high-quality, decoupled clothing and human body from monocular video.
- For the clothed human body in visible areas, we introduce a novel representation, hmSDF, which can accurately segment 3D clothing and body through easily obtainable 2D human parsing, without any 3D clothing priors.
- The reconstructed decoupled clothing and human body can be easily applied to animation production and clothing transfer applications, offering realistic and detailed geometric quality.

2. Related Work

Decoupled Representation of Clothed Human Body.

Most methods reconstruct the clothed human body geometry as a whole, including representations like mesh [2, 3], point clouds [34, 55], SDF [4, 6, 21, 37], and occupancy [46, 47]. These reconstruction methods maintain good detail in visible areas but are inconvenient for applications like animation and clothing transfer. A more effective approach is to independently represent clothing and the body, modeling them as decoupled, layered representations. GALA [27] and ClothCap [42] utilize 3D segmentation to obtain separate clothing and body meshes from 3D and 4D scans, respectively, a method constrained by the high cost of acquiring scan data. Neural-ABC [5] constructs a decoupled human body and clothing parametric model based on UDF (unsigned distance function) representation, but its re-

construction of details is limited. SCARF [9] reconstructs a mixed representation of a clothed human body from video sequences, but the NeRF [35] representation of clothing is limited in geometric effect. ReLoo [15] utilizes a double-layer SDF to reconstruct clothed human bodies. However, due to the watertight nature of SDF, it cannot directly reconstruct open meshes. Some other methods focus on reconstructing clothing [7, 8, 20, 45, 56], which allows decoupling by using SMPL [32] as the underlying body; however, the body often lacks detail in these approaches.

Reconstruction from Monocular Video. 3D clothed human body reconstruction from video inputs typically relies on motion and deformation cues to recover deformable 3D surfaces. Early works acquired actor-specific rigged templates [16, 17, 54] or used a parametric model as a prior [2]. Many efforts [9, 23, 39–41, 52], based on NeRF [35] and 3DGS [25], reconstruct animatable human avatars from dynamic videos, primarily focusing on rendering effects, but the quality of geometric reconstruction is not ideal. [14, 21, 51] reconstructed high-quality clothed human body geometries, but the clothing and underlying body are not separable. DGarment [28] and REC-MV [43] reconstructed dynamic clothing, excluding the body. Our method can reconstruct decoupled clothing and the underlying body while ensuring geometric quality.

3. Method

Given a monocular video containing N frames, which depicts a person in motion wearing clothing, as $\{I_t \mid t = 1, \dots, N\}$. D³-Human aims to reconstruct high-fidelity, decoupled, and spatio-temporally continuous clothing and underlying body meshes $\{G_t \mid t = 1, \dots, N\}$ without using 3D clothing template priors. With the goal of achieving maximum realism, the observed areas captured by the video, such as the exposed head and clothing, should be reconstructed with great detail; areas obscured, such as body parts covered by clothing or self-occluded by the body, should be reconstructed as plausibly as possible.

To achieve these objectives, We combine the flexibility of implicit representations with the robustness and rapid rendering capabilities of explicit representations to achieve the best results. Notably, the surface of a watertight clothed human can be segmented into clothing and body parts by closed curves. Therefore, we first utilize image and human parsing segmentation sequence information to reconstruct the separated clothing and body meshes in the visible areas. Then, with the help of SMPL [32], we complete the invisible body areas and generate decoupled clothing and body templates. Finally, normal information is additionally used to jointly optimize the clothing and body to enhance details.

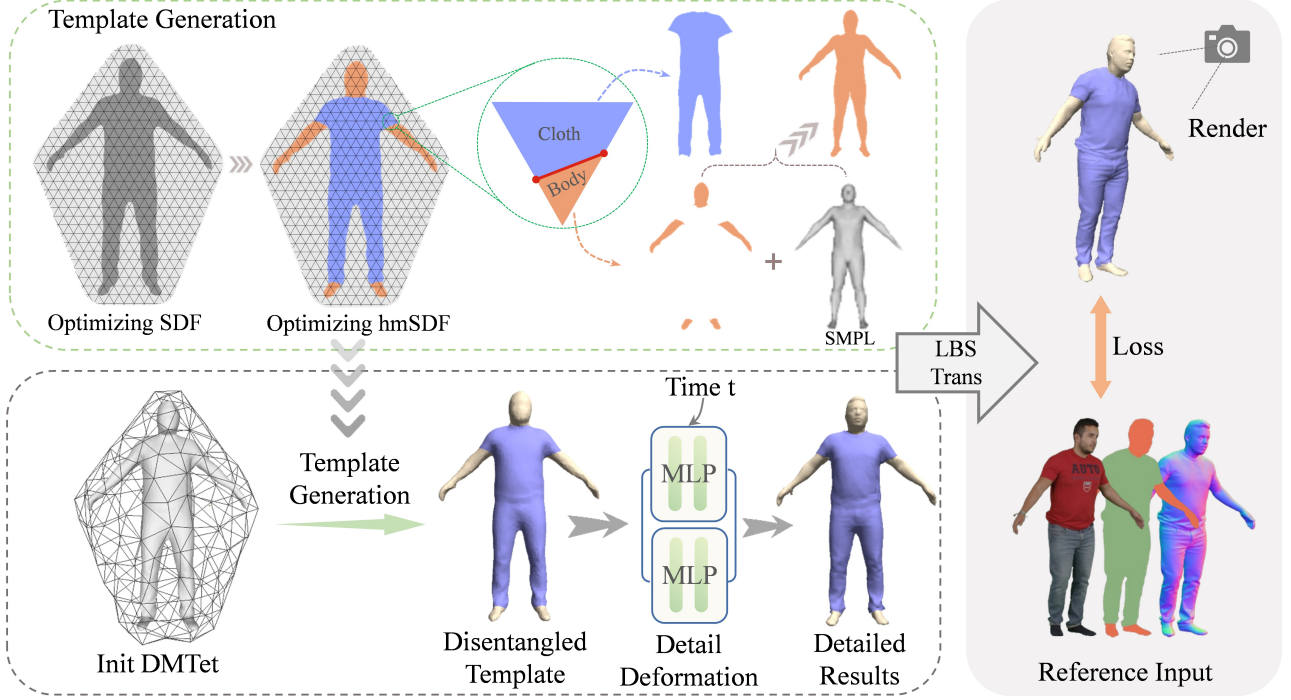


Figure 1. The algorithm of D³-Human composes two stages: template generation and detailed deformation. In the template generation stage, the watertight clothed human body is segmented via hmSDF and completed using SMPL, yielding disentangled clothing and body meshes. During the detailed deformation stage, two neural networks are utilized to model non-rigid deformations for clothing and body. Both stages employ LBS deformation from canonical to observed space and are rendered into 2D images for loss computation.

3.1. Overview

The proposed method can be divided into two stages: decoupled template generation and detailed deformation. Figure 1 shows the overall pipeline of our approach. The template is defined in the canonical pose space and deformed to the observed space using LBS.

In the template generation stage, the initialized DM Tet [50] is optimized into a watertight clothed human and then segmented into clothing and body regions using our proposed optimizable hmSDF. The regions fully occluded by clothing cannot be directly reconstructed from images, so we complete them using SMPL. After generating the decoupled template, we use two MLPs to model detailed deformations for each frame of the body and clothing meshes separately. During optimization, both the decoupled template and the meshes with detailed deformations are transformed to the observed space using forward LBS deformation and are supervised by images, normal maps, and parsing masks through a differentiable renderer.

3.2. HmSDF Representation

In this section, we define the clothing template G_c and the body template G_b in canonical space. Considering that the body parts obscured by clothing are invisible in the video,

we further divide the clothing and body into the visible body S_b , the invisible body U_b , and the visible clothing S_c . The visible body S_b is obtained via segmentation after reconstructing the whole clothed body, while the invisible body U_b is completed using the SMPL model. The visible and invisible bodies are merged to form the body template G_b .

The visible clothing and body are represented by a hybrid representations[50] that combines a tetrahedral mesh grid (V_T, T) and a neural implicit signed distance function $s_\eta(x)$, where $x \in V_T$ and $s_\eta(x)$ is a neural network with learnable weights η . The surface of $S_b \cup S_c$ can be represented by $S_\eta = \{x \in \mathbb{R}^3 \mid s_\eta(x; \eta) = 0\}$. The mesh can be extracted using methods following G-Shell [30].

Since the reconstructed watertight clothed human body $S_\eta = S_b \cup S_c$, we define a continuous and differentiable mapping $\nu : S_\eta \rightarrow \mathbb{R}$ on S_η to characterize whether a point belongs to S_b or S_c :

$$\nu(x) = \begin{cases} < 0, & \forall x \in \text{Interior}(S_b), \\ = 0, & \forall x \in \lambda, \\ > 0, & \forall x \in \text{Interior}(S_c), \end{cases}$$

where λ represents the boundary line between S_b and S_c . We denote ν as the human manifold signed distance fields, termed hmSDF. This differs from the definition of mSDF

in G-Shell, which only considers points located inside the open surface. In mSDF, regions > 0 are inside (valid), and < 0 are outside (invalid). In contrast, our approach considers points on both sides of the hmSDF (the visible clothing and body). In hmSDF, regions > 0 are clothing S_c (mSDF Case 1), and < 0 are body S_b (mSDF Case 3). There are no invalid regions exist in the watertight clothing-body surface.

3.3. Region Aggregation.

Ideally, an optimized hmSDF function ν should accurately segment the visible clothing S_c and body S_b . However, due to inaccuracies in the human parsing mask of each frame and inconsistencies between frames, inaccuracies may occur in the boundary line λ , as illustrated in Figure 2. When inaccuracies occur in the neighborhood of $\nu(x) = 0$, it causes the segmented regions S'_b and S'_c to contain holes and creates undesired segmented regions S''_b and S''_c . The small fragments contained in S''_b and S''_c are inevitably disconnected from other subgraphs.

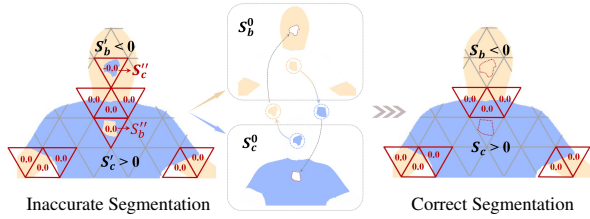


Figure 2. Schematic of region aggregation. For the correct segmentation results, S_b and S_c correctly segment the body and the cloth. For inaccurate segmentation results, S'_c should merge with S'_b , and S'_b should merge with S'_c .

We obtain the number of connected components for each category from the input image. Since the vertices of incorrectly classified connected components are usually fewer, the category of each subgraph, S'_b , S'_c , S''_b , and S''_c , can be determined by calculating the number of vertices in each connected component using the depth-first search method. The correct S_b and S_c can be obtained by aggregation as follows:

$$S_b = \text{merge}(S'_b, S''_b), \quad (1)$$

$$S_c = \text{merge}(S'_c, S''_c). \quad (2)$$

We demonstrate the process of aggregation of S_b and S_c in Algorithm 1.

3.4. Deformation Fields

Similar to previous methods [21, 43], we use the SMPL-based Linear Blend Skinning (LBS) method to model large deformations based on skeletal movements, and employ non-rigid deformation fields to model subtle deformations. However, a key difference is that clothing and body follow different motion rules. Therefore, we use two separate

Algorithm 1 Region Aggregation Algorithm

Input: The initialization segmentation S_b^0 and S_c^0 directly obtained from λ ; The correct number of subgraphs for S_b and S_c are α_1 and α_2 , respectively.

Calculate all connected subsets Q_b of S_b^0 ;

Calculate all connected subsets Q_c of S_c^0 ;

Sort Q_b based on the number of vertices;

Sort Q_c based on the number of vertices;

Extract S'_b and S''_b based on α_1 from Q_b ;

Extract S'_c and S''_c based on α_2 from Q_c ;

Obtain S_b by Equation 1 and filter out duplicate points.

Obtain S_c by Equation 2 and filter out duplicate points.

return S_b and S_c without holes and fragments.

non-rigid deformation fields to model the non-rigid deformations of clothing and body respectively.

Non-rigid Deformation. Due to its limited degrees of freedom, LBS deformation can only model large deformations and is unable to represent smaller details, such as the folds of clothing. Therefore, for detail deformation, we use two MLPs to model the non-rigid deformations of clothing and the human body. D is the MLP for non-rigid deformations:

$$x^t = D(x, h^t, E(x); \phi),$$

where x is the point in canonical space, x^t is the point after deformation in frame t , h^t is the latent code corresponding to frame t , $E(x)$ is the positional encoding, and ϕ are the network parameters that need to be optimized. For clothing and the body, the networks and parameters are independently separate.

LBS Deformation. Linear Blend Skinning (LBS) deformation models the transformation from canonical space to observed space based on skeletal deformation. Given the SMPL shape parameters β and pose parameters θ_t , the LBS deformation W can be written as:

$$G'(\beta, \theta_t) = W(D(x), \beta, \theta_t, \mathcal{W}(x)),$$

where $D(x)$ represents the non-rigid deformation of clothing and body, G' is the deformed mesh, and $\mathcal{W}(x)$ is the method for computing the skinning weights of x based on SMPL. We refer to some clothing simulation methods [11, 49] to calculate the skinning weights. For both clothing and body, we use a shared skinning deformation model.

3.5. Occlusion-Aware Differentiable Rendering

Differentiable rendering is used to render the geometry in the observed space to 2D, which allows for the calculation of loss with 2D supervision. Following some differentiable rendering methods [18, 30, 36], we utilize a differentiable rasterization approach to render the mesh. Compared to

volumetric rendering methods, rasterization-based rendering enables differentiable rendering for explicit meshes and offers better time and memory efficiency.

For clothed bodies, occlusion may occur between the body and clothing, which can lead to occlusion when rendering the visible areas of the clothing from the same viewpoint. Therefore, only rendering the clothing mesh to obtain a clothing mask may produce results inconsistent with the supervision signals, as illustrated in Figure 3. When rendering the clothing mesh alone, the obtained mask (d) cannot be used to correctly compute the loss with the GT mask (c). One approach is to render both the clothing and body meshes simultaneously and then extract the clothing region from the mask (e), allowing for correct loss computation with the GT mask (c). To solve this problem, we label the faces of both the clothing and body, rendering them simultaneously. We then use rasterization to generate occlusion-aware 2D labels, where the effective area of the body label is the body mask M_b , and the effective area of the clothing label is the clothing mask M_c .

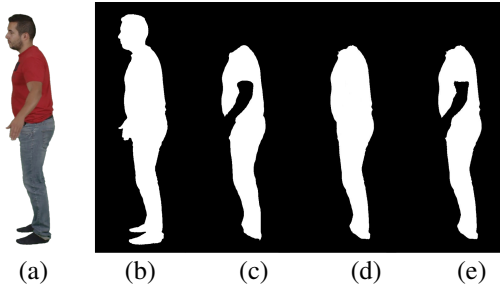


Figure 3. Demonstrating the differences between rendering of individual garments and rendering dressed human bodies when hands are occluded. (a) is the color image of the captured clothed human; (b) is the complete clothed body mask obtained from SAM2 [44]; (c) is the clothing mask obtained from SAM2; (d) is the mask obtained by rendering only the clothing mesh; (e) is the mask of the effective clothing area after rendering the complete clothed body mesh.

3.6. Training

3.6.1. Training Strategy

Our method consists of two stages: template generation and detail deformation optimization. In the template generation stage, we leverage hmSDF directly learn the clothing template from mask supervision, without relying on 3D clothing priors. In this stage, deformation is achieved solely through LBS, and parameters are optimized through RGB Loss, Mask Loss, Eikonal Loss, Encourage Hole Opening, and Regularize Holes. In the detail deformation stage, we introduce an additional Perceptual Normal Loss as a reconstruction term, while regularization is applied through Collision Penalty and Geometry Regularization to optimize non-rigid deformation.

3.6.2. Reconstruction Loss

We minimize the difference between the rendered result and the input image through the following objective:

RGB Loss. The MSE loss L_{mse} is calculated between the rendered RGB image and the supervision image. We calculate for all valid pixels P as following,

$$L_{\text{color}} = \frac{1}{|P|} \sum_{p \in P} \left(\mathbb{1}_b(p) \cdot L_{\text{mse}}(I_b, \hat{I}_b) + \mathbb{1}_c(p) \cdot L_{\text{mse}}(I_c, \hat{I}_c) \right), \quad (3)$$

where $\mathbb{1}(p)$ indicates the category to which the current pixel point p belongs. When $\mathbb{1}_b(p)$ is true, pixel p belongs to the body; when $\mathbb{1}_c(p)$ is true, pixel p belongs to the cloth. Both $\mathbb{1}_b(p)$ and $\mathbb{1}_c(p)$ can be true simultaneously, or only one may be true, depending on the mask used for supervision. I_b is the rendered RGB image of the body, I_c is the rendered RGB image of the clothing, $\hat{I}_b$ is the ground truth RGB image of the body, and $\hat{I}_c$ is the ground truth RGB image of the clothing.

Mask Loss. Although irrelevant backgrounds have already been removed in the RGB image supervision, the independently added mask loss can further constrain the accuracy of the edges, as,

$$L_{\text{mask}} = \frac{1}{|P|} \sum_{p \in P} \left(\mathbb{1}_b(p) \cdot L_{\text{mse}}(M_b, \hat{M}_b) + \mathbb{1}_c(p) \cdot L_{\text{mse}}(M_c, \hat{M}_c) \right), \quad (4)$$

where $\hat{M}_b$ is the ground truth mask of the body, and $\hat{M}_c$ is the ground truth mask of the clothing.

Perceptual Normal Loss. We obtain the normals of the image through Sapiens [26] as ground truth, to leverage the prior information trained on large-scale human body data. Rendered normals and supervision normals need to be normalized and aligned to the observation space. We use the perceptual loss [19, 24] to further enhance the effectiveness of the rendered normals.

$$L_{\text{per}} = \sum_i \|\phi_i(\mathcal{N}) - \phi_i(\hat{\mathcal{N}})\|^2, \quad (5)$$

where $\mathcal{N}$ is the rendered normal, and $\hat{\mathcal{N}}$ is the ground truth normal, and $\phi_i(*)$ denotes the activation of the i -th layer in the MobileNetV2 network [48].

3.6.3. Regularization Term

Eikonal Loss. To ensure a reasonable signed distance field, we add an Eikonal term [12] to the gradient g of the SDF value at each tetrahedral vertex when optimizing the SDF:

$$L_{\text{eik}} = \sum_{u \in V_T} (\|g_u\|_2 - 1)^2. \quad (6)$$

Encourage Hole Opening. With limited viewpoints, it is necessary to identify the opening positions using only image information. We encourage hmSDF openings by adopting a regularization term similar with [30] as,

$$L_{\text{hole}} = \sum_{u: \nu(u) \geq 0} L_{\text{huber}}(\nu(u)). \quad (7)$$

Regularize Holes. To avoid excessively large openings, we impose constraints on all points that are visible from the current viewpoint as

$$L_{\text{reg-hole}} = \sum_{u: \nu(u) = 0} L_{\text{huber}}(\nu(u) - \epsilon_1), \quad (8)$$

where ϵ_1 is a positive scalar.

Collision Penalty. This ensures that the garment does not penetrate the underlying body, inspired by [11, 49]. We implement it as

$$L_{\text{collision}} = \sum_{\text{vertices}} k_{\text{collision}} \max(\epsilon_2 - d(x), 0)^3. \quad (9)$$

In particular, when the distance between the two layers is too close, rendering can produce computational errors, so the value of ϵ_2 is set to 0.005.

Geometry Regularization. To ensure that the optimization is constrained, we encourage generating smooth deformed results. Inspired by Worchel et al. [53], we add normal consistency term $L_{\text{n-consist}}$ and Laplacian term $L_{\text{laplacian}}$.

4. Experiments

We conduct qualitative and quantitative experiments to demonstrate the effectiveness of D³-Human. For qualitative experiments, we use subjects from PeopleSnapshot [2] and SelfRecon [21]. For quantitative experiments, we use the synthetic dataset constructed by SelfRecon. It provides accurate ground truth for the meshes. Additionally, we perform ablation studies on the discussion of UDF and hmSDF, as well as the effectiveness of the perceptual normal loss, and demonstrate applications in clothing transfer and physics-based animation production.

4.1. Quantitative Evaluation.

Since there is no public real dataset available for evaluating the geometric quality of decoupled clothed human reconstruction from monocular video, we use four synthetic datasets provided by SelfRecon [21], each containing geometric ground truth and rendered videos. As these processed data are not open source from REC-MV [43], we manually labeled the split points and used the official tools to generate the feature lines. We also employ CLO3D [1] to separate clothing and body from the source data for quantitative evaluation. We report the Chamfer Distance (CD)

for clothing, body, and the complete clothed human of each method’s results. For methods that do not support decoupling, we only report the CD for the complete clothed human. We present the visualization of the quantitative comparison in Figure 4 and report the metrics in Table 1. The results show that our method achieves the best results in the metrics, provides detailed and accurate visual effects, and is capable of correctly decoupling the human body and clothing.

4.2. Qualitative Evaluation.

We compare our method to methods capable of reconstructing clothed human bodies from image sequences, using several sequences from the PeopleSnapshot [2] dataset and one sequence from SelfRecon [21] to include dress. For all methods, we extract consecutive frames from a complete rotation and present the comparison results for the first frame in Figure 7. The additional sequential results and discussion are presented in the supplementary materials.

As we can see, REC-MV [43] can accurately reconstruct clothing but lacks the level of detail achieved by our method. Due to the lack of further optimization for the body in REC-MV, directly using SMPL results in mesh penetration. BCNet [20] reconstructs clothing that can directly use SMPL as the underlying body; however, it only supports reconstruction with consistent clothing categories and lacks almost all detail. DELTA [10] enhances head details based on SCARF [9], and allows direct decoupled reconstruction of clothing and body. However, because it uses NeRF [35] representation for clothing, it cannot extract smooth geometry, leading to numerous artifacts in the clothing geometry. SelfRecon [21] uses the SDF representation to reconstruct clothed human bodies with correct shapes but lacking detail. GoMAvatar [52] employs a Gaussians-on-Mesh representation, resulting in relatively coarse meshes. Neither method can achieve decoupling of clothing from the body. Compared to these methods, our approach successfully decouples clothing from the body while maintaining a richer level of detail.

4.3. Ablation Study.

Implicit UDF or hmSDF? Several articles [5, 8, 30, 31] have demonstrated results by using the implicit Unsigned Distance Field (UDF) to represent clothing with undefined categories, leveraging dense supervision from meshes or multi-view images. However, we found that UDF struggles to produce robust results due to the limited supervision provided by single-view dynamic human reconstruction, as shown in Figure 5. The UDF reconstruction resulted in numerous small holes, a large hole in the abdominal area, and failed to create an opening at the cuffs. In contrast, hmSDF achieved an accurate garment shape.

Although UDF uses a network to model shapes and pos-

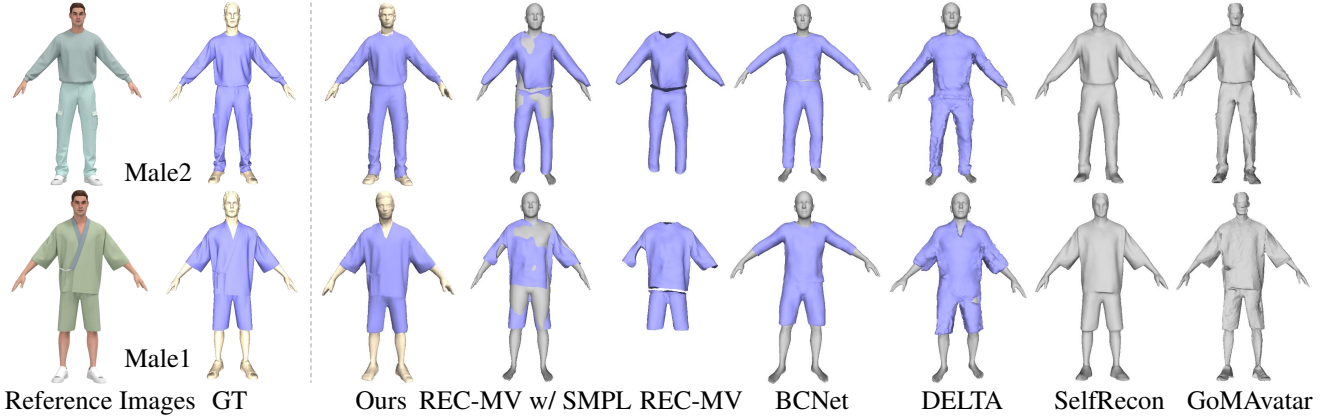


Figure 4. Quantitative comparison of the proposed method with REC-MV [43], BCNet [20], DELTA [10], SelfRecon [21], and GoMAvatar [52]. We use purple to visualize clothing that can be decoupled from the body. For REC-MV and BCNet, SMPL [32] was added as the body to show the complete reconstruction of the clothed human.

Table 1. Qualitative comparison across four synthetic sequences. We report the Chamfer Distance (CD) between the reconstructed surfaces (cm) and ground truth. For REC-MV, BCNet, DELTA and our method, we report the CD for clothing, body, and the full clothed human body, respectively. For SelfRecon and GoMAvatar, we only report the CD for the full clothed human body. The unit is e^{-3} . We highlight the **best** value and the **second best** value. We show Male1 and Male2 in Figure 4, and Female1 and Female3 in the supplementary materials.

Method	Female1			Female3			Male1			Male2		
	Clothing	Body	All	Clothing	Body	All	Clothing	Body	All	Clothing	Body	All
REC-MV [43]	1.416	1.789	1.148	0.930	2.082	1.461	0.614	1.945	0.619	0.693	1.201	0.616
BCNet [20]	1.685	10.252	5.561	4.571	10.112	5.681	2.589	6.236	4.802	2.007	4.109	2.853
DELTA [10]	2.177	0.973	1.388	2.173	0.820	0.915	1.327	1.498	1.702	1.884	1.073	1.132
SelfRecon [21]	-	-	3.420	-	-	2.249	-	-	1.310	-	-	1.454
GoMAvatar [52]	-	-	7.319	-	-	5.058	-	-	2.382	-	-	3.163
Ours	1.065	0.966	0.959	1.109	0.742	0.636	0.478	0.321	0.270	0.355	0.325	0.279

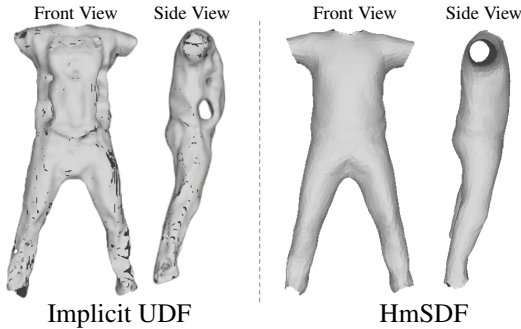


Figure 5. Ablation study on clothing reconstruction using implicit UDF [31] with deformation field [22] and hmSDF, applied to the male-3-casual data from the PeopleSnapshot [2] dataset.

sesses strong representational capabilities, it still has certain problems: (1) UDF is non-differentiable at the 0-level set. Although several solutions [13, 29, 31] have been proposed for mesh extraction and multi-view reconstruction problems, UDF remains sensitive near the 0-level set, and areas with poor supervision signals may lead to reconstruction failures. (2) The ability to extract surfaces from implicit UDF is limited. Surface extraction from implicit

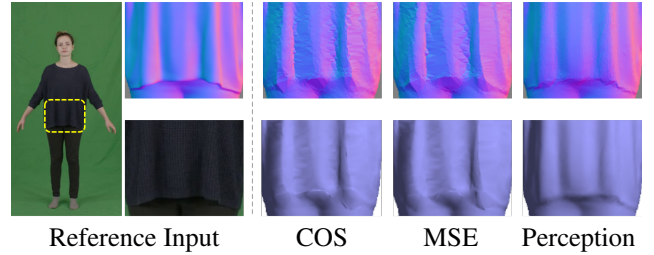
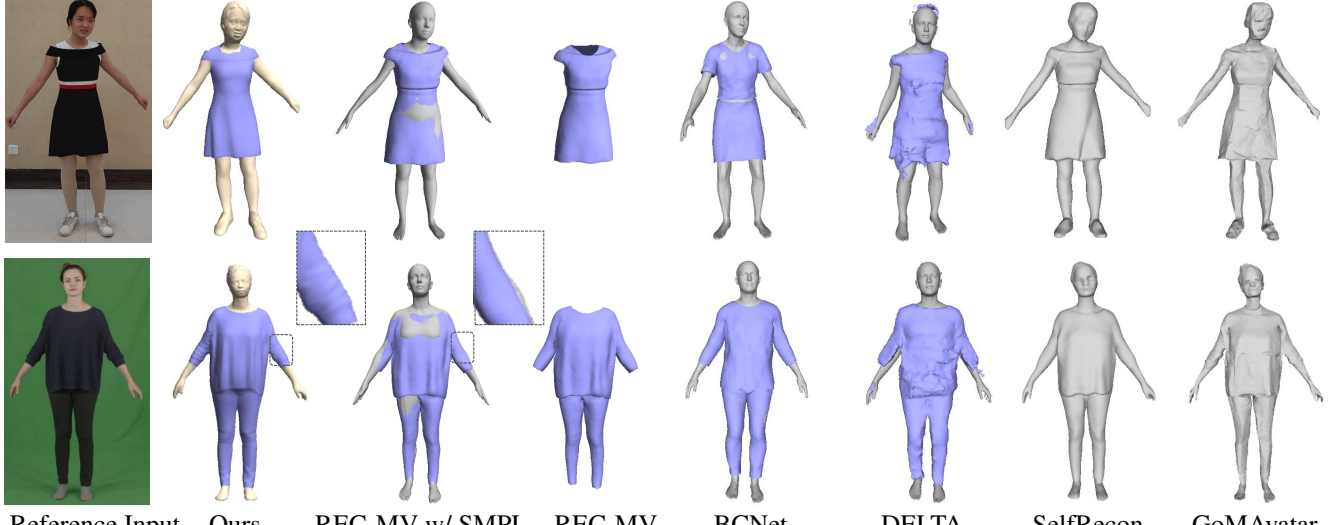


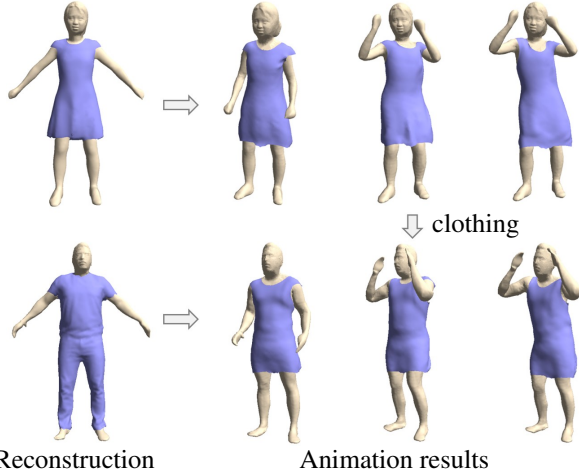
Figure 6. Ablation study on normal loss. The reference input includes normals and images. The top row shows the rendered normals, and the bottom row shows the rendered meshes.

UDF [13] is limited to manifold surfaces [33], and regions that are non-manifold in the implicit representation may fail to be extracted. (3) The strong representational capability of UDF reduces its noise resistance; for instance, the large hole in the abdominal area is caused by occlusion in the clothing mask. An example of occlusion is shown in Figure 3.

Perceptual Normal Loss. We attempt to remove the normal consistency loss and replace it with mean squared error and angular error. While the normal consistency loss focuses on feature consistency and can produce perceptu-



Reference Input Ours REC-MV w/ SMPL REC-MV BCNet DELTA SelfRecon GoMAvatar
Figure 7. Qualitative comparison. Comparison of our method with other methods on real image sequences. We visualize the first frame of each sequence. Compared to other methods, our results achieve better performance in reconstruction details and disentanglement capability.



Reconstruction Animation results
Figure 8. Decoupled reconstruction applications. Reconstructed clothing and body can be animated using physical simulation methods. Additionally, clothing can be easily swapped to create animations with different outfits.

ally consistent results, the mean squared error and angular error focus on pointwise features, which may result in less smooth results. We show the normal rendering results in Figure 6. Calculating the cosine or MSE loss between the rendered normals and the reference normals results in rough and noisy reconstructions. Using the perceptual normal loss, on the other hand, produces smoother results that preserve the features and details of the reference image.

4.4. Applications.

We demonstrate the clothing transfer after decoupled reconstruction and the application of physics-based animation.

The results are shown in Figure 8.

Clothing Transfer. Since our model is capable of reconstructing decoupled clothing and human bodies, clothing transfer can be achieved by separately reconstructing two clothed human figures and exchanging their garments.

Physics-based Animation. The reconstructed clothing and body geometry can be used with physics-based simulation methods to create more realistic animations. Compared to non-decoupled reconstructions [21], which make it difficult to establish motion between clothing and body, we use HOOD [11] to create more authentic clothing details.

5. Conclusion

We introduced D³-Human, a method that could reconstruct decoupled clothing and body directly from a short monocular video. By leveraging the robustness of explicit representations and the flexibility of implicit representations, D³-Human ensured the reconstruction of detailed features while maintaining the plausibility of body parts obscured by clothing. To achieve the separation of 3D clothing from the body, we proposed a novel representation called hmSDF defined on human body, which was able to obtain 3D segmentation using only 2D human parsing, without any 3D clothing priors. Thanks to this novel approach, we are able to achieve competitive reconstruction accuracy with much less computation time while ensuring the decoupling of clothing and body. The decoupled reconstruction results can be easily used for detailed animation production and clothing transfer. Our D³-Human can create high-quality and easily editable human geometry using only one camera, providing a technical foundation for the widespread adoption of many applications, such as highly editable digital avatar creation.

Acknowledgements. This research was supported by the National Natural Science Foundation of China (No.62441224, No.62272433). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of University of Science and Technology of China.

References

- [1] Clo3d, 2024. Accessed: 2024-11. 6
- [2] 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. 1, 2, 6, 7
- [3] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. *International Conference on Computer Vision*, 2019. 2
- [4] Yukang Cao, Kai Han, and Kwan-Yee K. Wong. Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [5] Honghu Chen, Yuxin Yao, and Juyong Zhang. Neural-abc: Neural parametric models for articulated body with clothes. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2, 6
- [6] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [7] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 2
- [8] Luca De Luigi, Ren Li, Benoit Guillard, Mathieu Salzmann, and Pascal Fua. DrapeNet: Garment Generation and Self-Supervised Draping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 6
- [9] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 1, 2, 6
- [10] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. *arXiv*, 2023. 6, 7
- [11] Artur Grigorev, Bernhard Thomaszewski, Michael J. Black, and Otmar Hilliges. HOOD: Hierarchical graphs for generalized modelling of clothing dynamics. 2023. 4, 6, 8
- [12] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020. 5
- [13] Benoit Guillard, Federico Stella, and Pascal Fua. Meshudf: Fast and differentiable meshing of unsigned distance field networks. In *European Conference on Computer Vision*, 2022. 2, 7
- [14] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2
- [15] Chen Guo, Tianjian Jiang, Manuel Kaufmann, Chengwei Zheng, Julien Valentin, Jie Song, and Otmar Hilliges. Reloo: Reconstructing humans dressed in loose garments from monocular video in the wild. In *European conference on computer vision (ECCV)*, 2024. 2
- [16] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. 2019. 1, 2
- [17] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 1, 2
- [18] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. *arXiv:2206.03380*, 2022. 4
- [19] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5
- [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 *European Conference on Computer Vision (ECCV)*, pages 18–35. Springer, 2020. 1, 2, 6, 7
- [21] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 4, 6, 7, 8
- [22] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. 2023. 7
- [23] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 1, 2
- [24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5
- [25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2
- [26] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 5
- [27] Taeksoo Kim, Byungjun Kim, Shunsuke Saito, and Hanbyul Joo. Gala: Generating animatable layered assets from a single scan, 2024. 2

- [28] Xiongzhen Li, Jinsong Zhang, Yu-Kun Lai, Jingyu Yang, and Kun Li. High-quality animatable dynamic garment reconstruction from monocular videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 2
- [29] Yu-Tao Liu, Li Wang, Jie Yang, Weikai Chen, Xiaoxu Meng, Bo Yang, and Lin Gao. Neudf: Learning neural unsigned distance fields with volume rendering. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 7
- [30] Zhen Liu, Yao Feng, Yuliang Xiu, Weiyang Liu, Liam Paull, Michael J. Black, and Bernhard Schölkopf. Ghost on the shell: An expressive representation of general 3d shapes. 2024. 2, 3, 4, 6
- [31] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Yuan Liu, Peng Wang, Christian Theobalt, Taku Komura, and Wenping Wang. Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies. *arXiv preprint arXiv:2211.14173*, 2022. 2, 6, 7
- [32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 2, 7
- [33] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 7
- [34] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 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, 6
- [36] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8280–8290, 2022. 4
- [37] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 1
- [39] P Paudel, A Khanal, A Chhatkuli, D Paudel, and J Tandukar. ihuman: Instant animatable digital humans from monocular videos. In *ECCV*, 2024. 1, 2
- [40] Bo Peng, Jun Hu, Jingtao Zhou, Xuan Gao, and Juyong Zhang. Intrinsicngp: Intrinsic coordinate based hash encoding for human nerf. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1
- [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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2
- [42] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4), 2017. Two first authors contributed equally. 2
- [43] Lingteng Qiu, Guanying Chen, Jiapeng Zhou, Mutian Xu, Junle Wang, and Xiaoguang Han. Recmv: Reconstructing 3d dynamic cloth from monocular videos. In *CVPR*, 2023. 1, 2, 4, 6, 7
- [44] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5
- [45] Li Ren, Benoit Guillard, and Pascal Fua. Isp: Multi-layered garment draping with implicit sewing patterns. *Advances in Neural Information Processing Systems*, 2023. 2
- [46] 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/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 2
- [47] 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
- [48] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5
- [49] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Snug: Self-supervised neural dynamic garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8140–8150, 2022. 4, 6
- [50] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3
- [51] Jeff Tan, Donglai Xiang, Shubham Tulsiani, Deva Ramanan, and Gengshan Yang. Dressrecon: Freeform 4d human reconstruction from monocular video, 2024. 2
- [52] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alex Schwing, and Shenlong Wang. GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh. In *CVPR*, 2024. 2, 6, 7
- [53] Markus Worchel, Rodrigo Diaz, Weiwen Hu, Oliver Schreer, Ingo Feldmann, and Peter Eisert. Multi-view mesh reconstruction with neural deferred shading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6187–6197, 2022. 6
- [54] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian

- Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.*, 37(2):27:1–27:15, 2018. [1](#), [2](#)
- [55] Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. Closet: Modeling clothed humans on continuous surface with explicit template decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. [2](#)
- [56] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3845–3854, 2022. [2](#)