High-Quality Animatable Dynamic Garment Reconstruction From Monocular Videos

Xiongzheng Li, Jinsong Zhang, Yu-Kun Lai, *Member, IEEE*, Jingyu Yang, *Senior Member, IEEE*, and Kun Li, *Member, IEEE*

***Abstract*— Much progress has been made in reconstructing garments from an image or a video. However, none of existing works meet the expectations of digitizing high-quality animatable dynamic garments that can be adjusted to various unseen poses. In this paper, we propose the first method to recover high-quality animatable dynamic garments from monocular videos without depending on scanned data. To generate reasonable deformations for various unseen poses, we propose a learnable garment deformation network that formulates the garment reconstruction task as a pose-driven deformation problem. To alleviate the ambiguity estimating 3D garments from monocular videos, we design a multi-hypothesis deformation module that learns spatial representations of multiple plausible deformations. Experimental results on several public datasets demonstrate that our method can reconstruct high-quality dynamic garments with coherent surface details, which can be easily animated under unseen poses. The code is available for research purposes at http://cic.tju.edu.cn/faculty/likun/projects/DGarment.**

***Index Terms*— High-quality, animatable, dynamic, monocular.**

I. Introduction

3D HUMAN digitization [1], [2], [3] is an active area in computer vision and graphics, which has a variety of applications in the fields of VR/AR [4], [5], fashion design [6] and virtual try-on [7], [8]. A fundamental challenge in digitizing humans is the modeling of high-quality animatable dynamic garments with realistic surface details, which can be adjusted to various poses. However, traditional methods require manual processes that are time-consuming even for an expert. Therefore, it is necessary to develop new methods that efficiently generate visually high-quality animatable dynamic 3D clothing without specialized knowledge.

Learning-based clothing reconstruction methods have been demonstrated to be feasible solutions to this problem. Early methods [9], [10], [11], [12], [13], [14], [15], [16] adopt a 3D scanner or a multi-view studio, but the high cost and large-scale setups prevent the widespread applications of such systems. For users, it is more convenient and cheaper to adopt a widely available RGB camera. Therefore, some works [17], [18], [19], [20], [21] attempt to reconstruct high-quality clothed humans from an RGB image or a monocular video. However, these methods use a single surface to represent both clothing and body, which fails to support applications such as virtual try-on. Layered representation with garment reconstruction [22], [23], [24], [25], [26] is more flexible and controllable, but related research works are relatively rare. Some methods [22], [23] adopt explicit parametric models trained on the Digital Wardrobes dataset [22], which can be adjusted to various unseen poses, but they fail to reconstruct garments with high-frequency surface details (*e.g.*, wrinkles). Other methods [24], [26] try to register explicit garment templates to implicit fields to improve reconstruction quality. However, this design leaves out the body pose, which makes it impossible to control or animate the garments flexibly. In addition, all the above methods not only rely on expensive data for training, but are also bounded by domain gaps and cannot generalize well to the inputs outside the domain of the training dataset. Most importantly, none of these works meet the expectations of digitizing high-quality animatable dynamic garments that can be adjusted to various unseen poses.

Therefore, our goal is to reconstruct high-quality animatable dynamic garments from monocular videos. There are major challenges that need to be overcome to achieve this: 1) a large amount of scanned data is needed for supervision, which tends to result in domain gaps and limited performance for unseen data; 2) the absence of strong and efficient human priors increases the difficulty of estimating dynamic and reasonably wrinkled clothing directly from monocular videos; 3) recovering dynamic 3D clothes from monocular videos is a highly uncertain and inherently ill-posed problem due to the depth ambiguity.

In this paper, we propose a novel weakly supervised framework to reconstruct high-quality animatable dynamic garments from monocular videos, aiming to eliminate the need to simulate or scan hundreds or even thousands of human sequences. By applying weakly supervised training, we greatly reduce the required time of both data preparation and model deployment. To the best of our knowledge, our method is the first work to reconstruct high-quality animatable dynamic garments from a single RGB camera without depending on scanned data.

To handle dynamic garment deformation from monocular videos, we propose a learnable garment deformation

Manuscript received 17 May 2023; revised 25 September 2023; accepted 26 October 2023. Date of publication 3 November 2023; date of current version 6 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62171317 and Grant 62122058. This article was recommended by Associate Editor H.-C. Shih. *(Corresponding author: Kun Li.)*

Xiongzheng Li, Jinsong Zhang, and Kun Li are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China (e-mail: lik@tju.edu.cn).

Yu-Kun Lai is with the School of Computer Science and Informatics, Cardiff University, CF24 4AG Cardiff, U.K.

Jingyu Yang is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCSVT.2023.3329972.

Digital Object Identifier 10.1109/TCSVT.2023.3329972

1051-8215 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

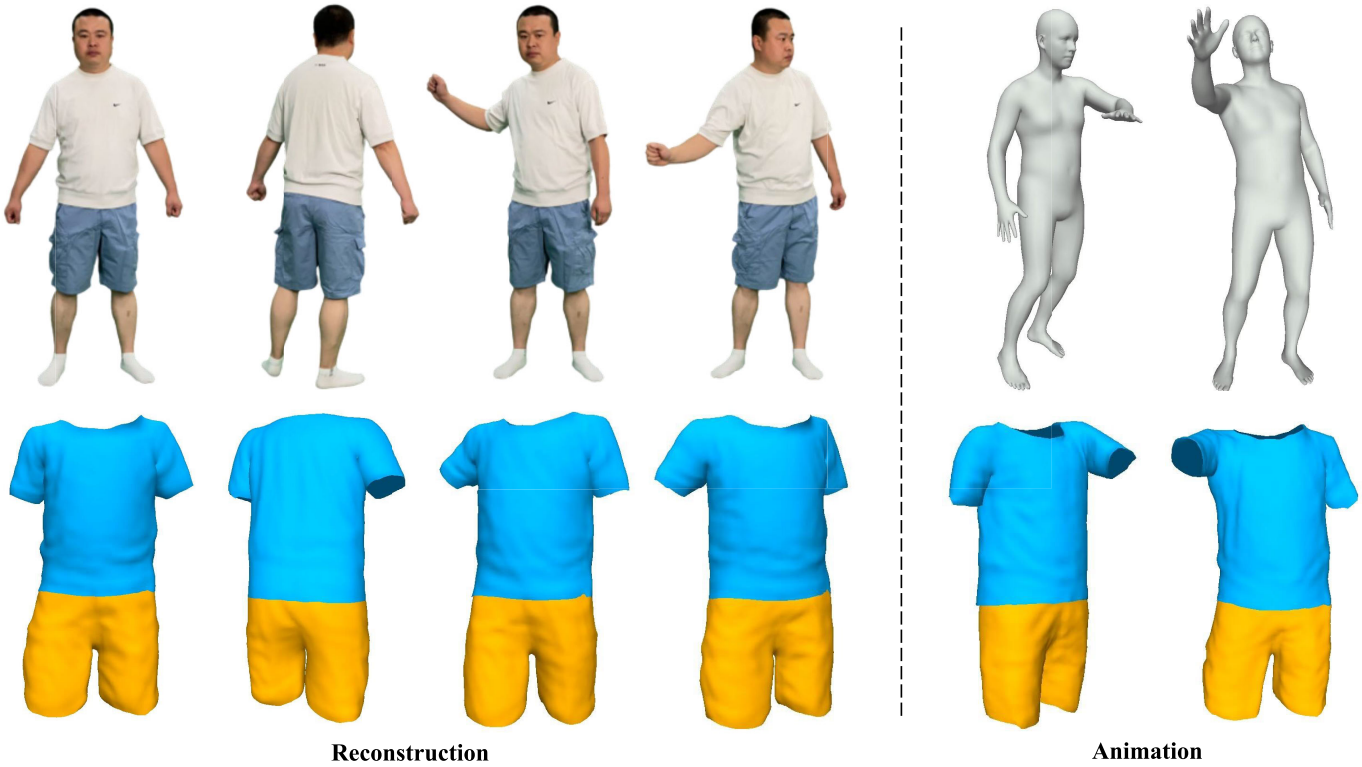

Fig. 1. Given a video of a person, our method can reconstruct high-quality and animatable garments, which enables new deformations for various unseen poses to be generated.

network that formulates the garment reconstruction task as a pose-driven deformation problem. In particular, we utilize human body priors [27] to guide the deformation of the spatial points of garments, which makes the garment deformation more controllable and enables our model to generate reasonable deformations for various unseen poses. To alleviate the ambiguity resulted from estimating 3D garments from monocular videos, we design a simple but effective multi-hypothesis displacement module that learns spatial representations of multiple plausible deformation. We observe that it is more reasonable to conduct multi-hypothesis estimation to obtain garment deformation than direct regression, especially for monocular camera settings, as this way can enrich the diversity of features and produce a better integration for the final 3D garments. The prior works [22], [23], [25] focus on the geometry of the clothes and do not attempt to recover the garment textures, which limits their application scenarios. Therefore, we design a neural texture network to generate high-fidelity textures consistent with the image. Experimental results on several public datasets demonstrate that our method can reconstruct high-quality dynamic garments with coherent surface details, which can be easily animated under unseen poses. An example is given in Fig. 1. *The code is available for research purposes at http://cic.tju.edu.cn/faculty/likun/projects/DGarment.*

Our main contributions can be summarized as follows:

- We design a weakly supervised framework to recover high-quality dynamic animatable garments from a monocular video without depending on scanned data. To the best of our knowledge, no other work meets the expectations of digitizing high-quality garments that can be adjusted to various unseen poses.
- We propose a learnable garment deformation network that formulates the garment reconstruction task as a pose-driven deformation problem based on human body priors. This enables our model to generate reasonable deformations for various unseen poses.
- We propose a simple but effective multi-hypothesis displacement module that learns spatial representations of multiple plausible deformations. In this way, we can alleviate the ambiguity brought by estimating 3D garments based on monocular videos.

II. Related Work

A. Clothed Human Reconstruction

Clothed human reconstruction is inevitably challenging due to complex geometric deformations under various body shapes and poses. Some methods [18], [28], [29], [30], [31], [32], [33] explicitly model 3D humans based on parametric models like SMPL [27], and as a result may fail to accurately recover 3D geometry. Zhu et al. [34] combine a parametric model with flexible free-form deformation by leveraging a hierarchical mesh deformation framework on top of the SMPL model [27] to refine the 3D geometry. These methods predict more robust results, but fail to reconstruct garments with high-frequency surface details. Contrary to parametric-model-based methods, non-parametric approaches directly predict the 3D representation from an RGB image or a monocular video. Zheng et al. [35] propose an image-guided volume-to-volume translation framework fused with

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

image features to reproduce accurate surface geometry. However, this representation requires intensive memory and has low resolution. To avoid high memory requirements, implicit function [19], [36] representations are proposed for clothed human reconstruction. Saito et al. [19] propose a pixel-aligned implicit function representation called PIFu for high-quality mesh reconstructions with fine geometry details (*e.g.*, clothing wrinkles) from images. However, PIFu [19] and its variants [37], [38], [39], [40], [41], [42], [43] may generate implausible results such as broken legs. Feng et al. [20] propose a new 3D representation, FOF (Fourier Occupancy Field), for monocular real-time human reconstruction. Nonetheless, FOF cannot represent very thin geometry restricted by the use of low-frequency terms of Fourier series. Recently, inspired by the success of neural rendering methods in scene reconstruction [44], [45], various methods [46], [47], [48], [49], [50], [51] recover 3D clothed humans directly from multi-view or monocular RGB videos. Although these approaches demonstrate impressive performance, they fail to support applications such as virtual try-on, because they use a single surface to represent both clothing and body.

B. Garment Reconstruction

In comparison to clothed human reconstruction using a single surface representation for both body and clothing, treating clothing as separate layers on top of the human body [22], [23], [24], [25], [26], [52], [53], [54], [55], [56] allows controlling or animating the garments flexibly and can be exploited in a range of applications. Some methods [22], [23], [24], [25], [26], [52], [53] address the challenging problem of garment reconstruction from a single-view image. Bhatnagar et al. [22] propose the first method to predict clothing layered on top of the SMPL [27] model from a few frames of a video trained on the Digital Wardrobes dataset. Jiang et al. [23] split clothing vertices off the body mesh and train a specific network to estimate the garment skinning weights, which enables the joint reconstruction of body and loose garment. SMPLicit [25] is another approach that builds a generative model which embeds 3D clothes as latent codes to represent clothing styles and shapes. As a further attempt, Moon et al. [52] propose Clothwild based on SMPLicit [25] to produce robust results from in-the-wild images. Although these methods regard clothing and human body as independent layers, they fail to recover high-frequency garment geometry.

Unlike previous works, to reconstruct high-quality garment geometry, Deep Fashion3D [24] uses an implicit Occupancy Network [57] to model fine geometric details on garment surfaces. Zhu et al. [26] extend this idea by proposing a novel geometry inference network ReEF, which registers an explicit garment template to a pixel-aligned implicit field through progressive stages including template initialization, boundary alignment and shape fitting. Zhao et al. [53] utilize the predicted 3D anchor points to learn an unsigned distance function, which enables the handling of open garment surfaces with complex topology. However, these methods cannot deal with dynamic clothing, thus they are not suitable for dynamic garment reconstruction.

Other methods [9], [58], [59], [60], [61] try to reconstruct dynamic clothing from video. Garment Avatar [58] proposes a multi-view patterned clothing tracking algorithm capable of capturing deformations with high accuracy. Li et al. [9] propose a method for learning physically-aware clothing deformations from monocular videos, but their method relies on an individual-dependent 3D template mesh [59]. SCARF [60] combines the strengths of body mesh models (SMPL-X [62]) with the flexibility of NeRFs [45], but the geometry of clothing is sometimes noisy due to the limited 3D geometry quality for NeRF reconstruction. REC-MV [61] introduces a method to jointly optimize the explicit feature curves and the implicit signed distance field (SDF) of the garments to produce high-quality dynamic garment surfaces. These solutions show their strength in reconstructing high-fidelity layered representations with garments that remain in consensus with the input person. However, this design leaves out the body pose, which makes it impossible to control or animate garments flexibly.

In this paper, we design a weakly supervised framework to recover high-quality dynamic garments from a monocular video without depending on scanned data. In the meanwhile, we propose a learnable garment deformation network which enables our model to generate reasonable deformations for various unseen poses.

III. Method

Our goal is to reconstruct high-quality animatable dynamic garments from a monocular video, which effectively enables personalized clothing animation. Previous works not only rely on expensive data, but are also bounded by domain gaps and cannot generalize well to inputs outside the domain of the training dataset. Therefore, we propose to reconstruct clothes in a weakly supervised manner, thus addressing the main drawbacks of previous works in terms of cost. Given a monocular video which consists of a clothed human under random poses, we first extract human-centric information such as segmentation maps and normal maps [63], [64], [65], [66] to help obtain consistent geometry details with the input video (Sec. III-A). To enable the generation of animatable dynamic garments for various unseen poses, we propose a learnable garment deformation network based on human body priors which formulates the garment reconstruction task as a pose-driven deformation problem (Sec. III-B). In addition, different from previous works which estimate a unique displacement vector for each garment vertex, our method leverages a multi-hypothesis deformation module to alleviate the depth ambiguity and provide integrated deformations for the final reconstructed garments (Sec. III-C). The overview of our method is illustrated in Fig. 2.

A. Human-Centric Information Extracting

To get rid of costly data preparation, we design a weakly supervised framework to recover high-quality dynamic garments from monocular videos, and we extract human-centric information which is helpful to obtain consistent geometry details with the input. Specifically, given a monocular video $I = \{\mathbf{I}_0, \ldots, \mathbf{I}_{n-1}\}$, where n is the number of frames, we first

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

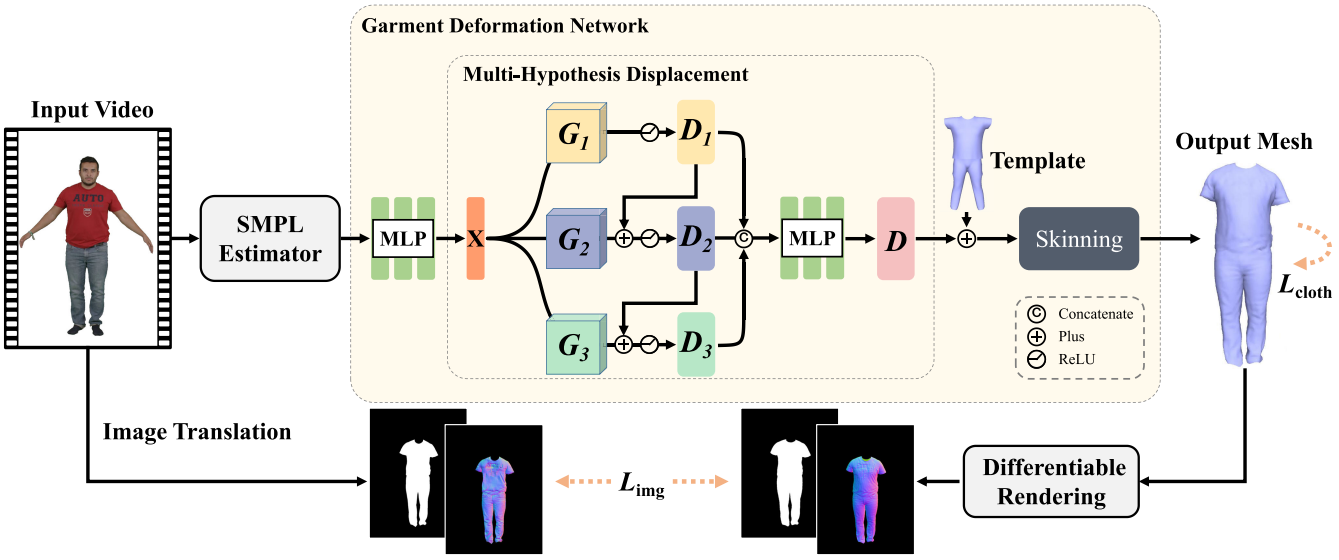

Fig. 2. Overview of our method. At the core of our method lies a learnable garment deformation network that predicts reasonable deformations for the input video. For each frame, we first design an MLP to obtain a high-level embedding X based on the SMPL pose and define three learnable matrices G_1, G_2, G_3 to get three deformations D_1, D_2, D_3. Then, we connect them as the input of an MLP that outputs the final deformation D. This design can enrich the diversity of features and help produce aggregated displacements for more reasonable garment deformation. These displacements are added to the garment template, which is then skinned along the body according to pose parameters θ and blend weights $\mathcal{W}_c$ to produce the final result. We train the network using L_{img} and L_{cloth} loss function in a weakly supervised manner, which removes the need for ground-truth data.

use a state-of-the-art human pose estimation method [67] to estimate the pose parameters $\theta \in \mathbb{R}^{72}$ and the shape parameters $\beta \in \mathbb{R}^{10}$ of a SMPL human body model [27], as well as the weak-perspective camera parameters $c \in \mathbb{R}^3$ for each frame. The pose parameters θ and the shape parameters β represent 3D rotations of human body joints and PCA (Principal Component Analysis) coefficients of T-posed body shape space, respectively. Second, we obtain the binary masks $Ms = \{\mathbf{Ms}_0, \ldots, \mathbf{Ms}_{n-1}\}$ of the input I using a robust human parsing method PGN [68]. Note that the output of PGN is a set of segmentation masks, where each pixel corresponds to a human body part or clothing type. We remove the masks of the human body, leaving only the ones of the clothing, and transform segmentation masks to binary masks. Third, we use PIFuHD [37] to estimate the image normals $Nor = \{\mathbf{Nor}_0, \ldots, \mathbf{Nor}_{n-1}\}$ of the input I and multiply them by the binary masks S to get the normal map of the garment. Finally, we obtain the smooth garment template $\mathbf{T}$ of the first frame under T-pose based on the work of Jiang et al. [23]. The information above helps reduce the complexity inherent to our garment reconstruction. Our template supports six garment categories, including upper garment, pants, and skirts with short and long templates for each type.

B. Garment Deformation Network

The absence of strong and efficient human priors increases the difficulty of estimating dynamic and reasonably wrinkled clothing directly from a monocular video. Different from previous works which extract features from images to generate clothing deformations, we observe that the garment deformation is caused by changes in pose. Therefore, we design a garment deformation network which enables our model to generate reasonable deformations for various unseen poses. To achieve this, we use a parametric SMPL model [27] to guide the deformation of the spatial points of garments [69], which enables explicit transformation from template space to current posed space. With the SMPL model, we can map the shape parameters β and the pose parameters θ to a body mesh $\mathbf{M}_b$. The mapping can be summarized as:

$$\begin{aligned} \mathbf{M}_b(\beta, \theta) &= W_b(\mathbf{T}_b(\beta, \theta), \mathbf{J}(\beta), \theta, \mathcal{W}_b), \\ \mathbf{T}_b(\beta, \theta) &= \mathbf{B} + \mathbf{B}_s(\beta) + \mathbf{B}_p(\theta), \end{aligned} \tag{1}$$

where W_b is the linear blend skinning function of the human body, $\mathbf{J}(\beta)$ is the SMPL body's skeleton, and $\mathcal{W}_b$ is the blend weights of each vertex of SMPL. $\mathbf{B}_s(\beta)$ and $\mathbf{B}_p(\theta)$ are the pose blendshape and shape blendshape, respectively. As most clothes follow the deformation of the body, we share garment pose parameters θ with SMPL and use SMPL's skeleton $\mathbf{J}(\beta)$ as the binding skeleton of the garment. In this way, we define our cloth mesh M_c as follows:

$$\begin{aligned} \mathbf{M}_c(\beta, \theta) &= W_c(\mathbf{T}_c(\theta), \mathbf{J}(\beta), \theta, \mathcal{W}_c), \\ \mathbf{T}_c(\theta) &= \mathbf{T} + D_\theta, \end{aligned} \tag{2}$$

where W_c is the linear blend skinning function of the garment, $\mathcal{W}_c$ is the blend weights of each vertex of the garment, $\mathbf{T}$ is the smooth garment template and D_θ is the high-frequency displacement over the garment template.

For the pose θ of each frame, we design a four-layer Multi-Layer Perceptron (MLP) with ReLU activation function to obtain a high-level embedding X, and further obtain the garment vertex deformations D_θ with a learnable matrix $G \in \mathbb{R}^{x \times N \times 3}$ (where x is the dimensionality of the high-level embedding and N is the number of vertices of the garment mesh). This non-linear mapping from θ to D_θ allows modeling high-frequency details, such as wrinkles caused by different poses, which are beyond the representation ability of the linear

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

model. For each vertex on the garment template, instead of directly using the skinning weights of SMPL, we assign its blend weights equal to those of the closest body vertex and allow the blend weights to be optimized during training to make the garment mesh independent from the SMPL. Our garment deformation network can reconstruct pose-dependent garments, which enables the generation of reasonable deformations for various unseen poses.

C. Multi-Hypothesis Displacement

Recovering 3D clothes from monocular videos is a highly uncertain and inherently ill-posed problem. We propose a multi-hypothesis displacement module that learns spatial representations of multiple plausible deformations in the learnable garment deformation network. Since each pixel of the image corresponds to innumerable points in the 3D space, it is difficult to specify a unique 3D point corresponding to a given pixel. To alleviate the depth ambiguity brought by estimating 3D garments from monocular video, different from previous works which estimate a unique displacement vector for each garment vertex, we design a cascaded architecture to generate multiple displacements using the high-level pose embedding X. More specifically, we first define three learnable matrices to get three deformations and encourage gradient propagation through residual connections. Then, we connect the three hypothetical deformations as the input of an MLP that outputs the final deformation. These procedures can be formulated as:

$$\begin{aligned} X &= \sigma(\mathcal{F}_{\text{MLP}_1}(\theta)), \\ D_1 &= \sigma(X \cdot G_1 + b_1), \\ D_2 &= \sigma(D_1 + X \cdot G_2 + b_2), \\ D_3 &= \sigma(D_2 + X \cdot G_3 + b_3), \\ D_\theta &= \mathcal{F}_{\text{MLP}_2}(D_1, D_2, D_3), \end{aligned} \tag{3}$$

where X is the high-level embedding mentioned in Sec. III-B, σ denotes the ReLU activation function, G_* and b_* are the learnable matrices and bias terms respectively, and $\mathcal{F}_{\text{MLP}_*}$ represents Multi-Layer Perceptron. For simplicity, we use * to represent an arbitrary subscript. With this design, our model can first predict multiple displacements, which can enrich the diversity of features, and then aggregate them to produce more reasonable displacements for the 3D garments. Finally, these displacements are added to the garment template to obtain the result in T-pose, which is then skinned along the body according to pose parameters θ and blend weights $\mathcal{W}_c$ to produce the final result.

D. Loss Function

The loss function of our weakly supervised network includes the constraints from the image and the geometric constraints of the clothes, which not only produces image-consistent details, but also keeps the garment stable. The overall loss function is

$$L = L_{\text{img}} + L_{\text{cloth}}. \tag{4}$$

1) Image Loss: To generate garment geometry and shape that are consistent with the input, we regularize the shape of clothing by projecting it onto an image, and compute the loss with the target mask $\mathbf{S}_i$ and we utilize the predicted normal map to further refine the geometry shape. We define the following image loss:

$$\begin{aligned} &L_{\text{img}} \\ &= \lambda_{\text{mask}} ||\mathcal{F}_{\text{mask}}(M_i, c) - \mathbf{Ms}_i||_2 \\ &\quad + \lambda_{\text{normal}} ||\mathcal{F}_{\text{VGG}}(\mathcal{F}_{\text{normal}}(M_i, c)) - \mathcal{F}_{\text{VGG}}(\mathbf{Ms}_i \cdot \mathbf{Nor}_i)||_2, \end{aligned} \tag{5}$$

where λ_{mask} and λ_{normal} are the weights that balance the contributions of individual loss terms. $\mathcal{F}_{\text{mask}}$ is a differentiable renderer [70] that renders the mask of garment mesh M_i corresponding to the i-th frame, given the camera parameters c. $\mathcal{F}_{\text{normal}}$ outputs the normal map in a similar way to $\mathcal{F}_{\text{mask}}$, $\mathbf{Ms}_i \cdot \mathbf{Nor}_i$ is the normal map of the garment as mentioned in Sec.III-A, and $\mathcal{F}_{\text{VGG}}$ is the VGG-16 network used to extract image features to help measure their similarity.

2) Clothing Loss: Using only the image loss is inclined to produce unstable results. Thus another clothing loss term is added to enhance stability of the reconstructed garments:

$$\begin{aligned} L_{\text{cloth}} &= \lambda_{\text{edge}} ||E - E_T||_2^2 + \lambda_{\text{face}} ||\Delta(N_F)||_2^2 \\ &\quad + \lambda_{\text{angle}} ||\Theta||_2^2 + \lambda_{\text{collision}} \sum_{j=0}^{V_b} \max(\varepsilon - d_j \cdot n_j, 0)^2. \end{aligned} \tag{6}$$

E is the predicted edge lengths, E_T is the edge lengths on the template garment $\mathbf{T}$, N_F is the face normals, $\Delta(\cdot)$ is the Laplace-Beltrami operator, and Θ is the dihedral angle between faces. λ_{edge}, λ_{face}, λ_{angle} and $\lambda_{\text{collision}}$ are the balancing weights. where d_j is the vector going from the j-th vertex of the body vertices V_b to the nearest vertex of the garment, n_j is the normal of the j-th body vertex, ε is a small positive threshold. On the one hand, inspired by [71] and [72], the first three terms of L_{cloth} ensure the clothing is not excessively stretched or compressed and enforces locally smooth surfaces. On the other hand, the last item of L_{cloth} is used to handle the collision between the clothes and the body.

E. Implementation Details

Our model is implemented using Tensorflow, and we train our model for 10 epochs with a batch size of 8 using the Adam optimizer [73] with a learning rate of 1×10^{-4}. The embedding dimensions of the MLP used to obtain the high-level embedding are set to 256, 256, 512 and 512, respectively, and the learnable matrix is initialized using the truncated normal distribution. We choose the weights of the individual losses with $\lambda_{\text{mask}} = 500$, $\lambda_{\text{normal}} = 1500$, $\lambda_{\text{edge}} = 100$, $\lambda_{\text{face}} = 2000$, $\lambda_{\text{angle}} = 1$ and $\lambda_{\text{collision}} = 100$. For a video of about 600 frames in length with a resolution of 512×512, we train our model with a Titan X GPU in half an hour.

IV. Neural Texture Generation

The prior works [22], [23], [25] focus on the geometry of the clothes and do not attempt to recover the garment textures, which limits the application scenarios. Texture is extremely

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

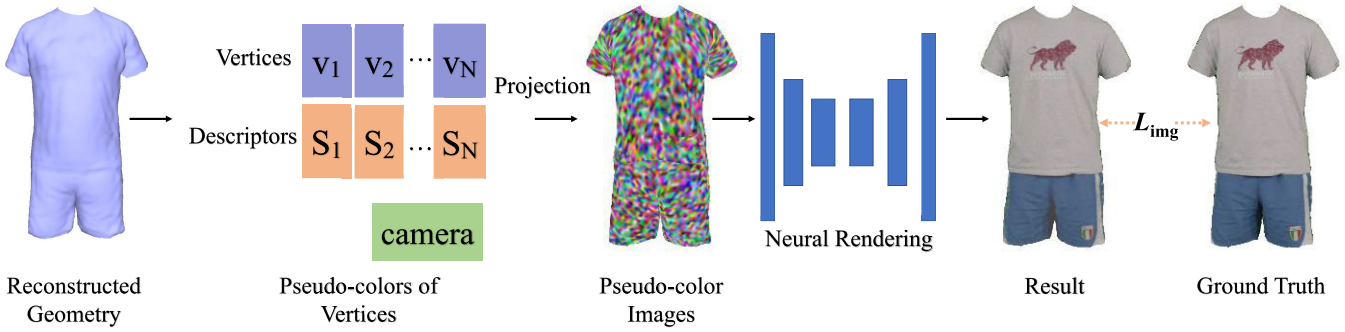

Fig. 3. An overview of our neural rendering pipeline. Given the reconstructed mesh with the descriptors and the camera, we first project the mesh onto the image plane, using descriptors as pseudo-colors. We then use the rendering network to transform the pseudo-color images into a photo-realistic RGB image.

complex: it resides in high-dimensional space and is difficult to represent. Therefore, to cope with the complexity of textures, we propose a neural texture network to obtain photo-realistic results. As different garment meshes have different topologies, it is computationally expensive to generate a UV map every time. Inspired by [75] and [76], our main idea is to combine the point-based graphics and neural rendering. Below, we will explain the details of our method. An overview of our neural rendering pipeline is illustrated in Figure 3.

Based on the multi-hypothesis displacement module, we get the relatively accurate geometry of clothing mesh M_c, which allows the neural texture network to focus on texture information. We first attach descriptors $S = \{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$ which serve as pseudo-colors, to the garment mesh vertices $V = \{\mathbf{V}_1, \ldots, \mathbf{V}_N\}$. We first project the mesh onto the image plane to obtain pseudo-color image R_{img}, then use the neural texture network to transform the pseudo-color image R_{img} into a photo-realistic RGB image I_{img}. Specifically, given the pseudo-color image and the ground truth image, we adpot a UNet-based neural texture network to map the initial mesh projections to the final output image. The neural texture network consists of 8 blocks of downsampling and 8 blocks of upsampling convolutional layers. Each downsampling block consists of a convolution layer with BatchNorm operations followed by ReLU activations; each upsampling block consists of a transposed convolution layer with BatchNorm operations followed by ReLU activations.

Using the ground-truth image I_{gt}, we optimize our neural texture network by minimizing the differences between the rendered image I_{img} and ground-truth RGB image I_{gt}. to obtain higher quality results, we adopt a two-stage training strategy. In the first stage, we optimize the descriptor to obtain a better initial value for the second stage. Specifically, during the first stage, we train the model using the Adam optimizer with a learning rate of 1×10^{-4} and the batch size of 4 for 25 epochs by minimizing the perceptual loss between pseudo-color image and ground-truth image:

$$L_{\text{pse}} = ||\mathcal{F}_{\text{VGG}}(R_{img}) - \mathcal{F}_{\text{VGG}}(I_{gt})||_2, \tag{7}$$

where $\mathcal{F}_{\text{VGG}}$ is the image features extracted from the VGG-16 network which is used to ensure the perceptual similarity.

During the second stage, we train the model for 25 epochs using the Adam optimizer with a learning rate of 1×10^{-4} which is decayed by a factor of 0.5 every 10 epochs:

$$L_{\text{render}} = ||\mathcal{F}_{\text{VGG}}(I_{gt}) - \mathcal{F}_{\text{VGG}}(I_{img})||_2 + \lambda_{\text{render}}||I_{gt} - I_{img}||_1, \tag{8}$$

where λ_{render} is the balancing weight and is set to 100 in our experiments. The overall training time is around 1.5 hours with a Titan X GPU.

V. Experiments

A. Datasets

To demonstrate the effectiveness of our proposed method, we conduct experiments on four different datasets: People-Snapshot [18], CAPE [74], IPER [77] and our captured data. People-Snapshot [18], IPER and our captured data contain different monocular RGB videos captured in real-world scenes, where subjects turn around with a rough A-pose in front of an RGB camera. In addition, IPER and our captured data also contain videos of the same person with random motions. CAPE [74] is a dynamic dataset of clothed humans which provides raw scans of 4 subjects performing simple motions. These four datasets are used to evaluate the quality of the 3D reconstructions, IPER and our captured data are also used to show the results of garment animation. The SMPL parameters provided by CAPE [74] and People-Snapshot [18] are used. For the input video, 80% is used for training (Reconstruction) and 20% is used for testing (Animation).

B. Comparison

We compare our method against the state-of-the-art garment reconstruction methods that release the codes: Multi-Garment Net (MGN) [22], BCNet [23], and SMPLicit [25], both qualitatively and quantitatively. Note that these methods all apply supervised learning, either using 3D scans or synthetic datasets to train the models, while we propose to reconstruct clothes in a weakly supervised manner without 3D supervision.

1) Qualitative Comparison: In Fig. 4, we show the visual results of the same person in three different poses. It can be seen that for different poses, our reconstruction method produces different deformations consistent with the image, while MGN [22] and SMPLicit [25] can only produce smooth results. BCNet [23] can generate some details, but not as rich as ours. In Fig. 5, since the person maintains a rough A-pose

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

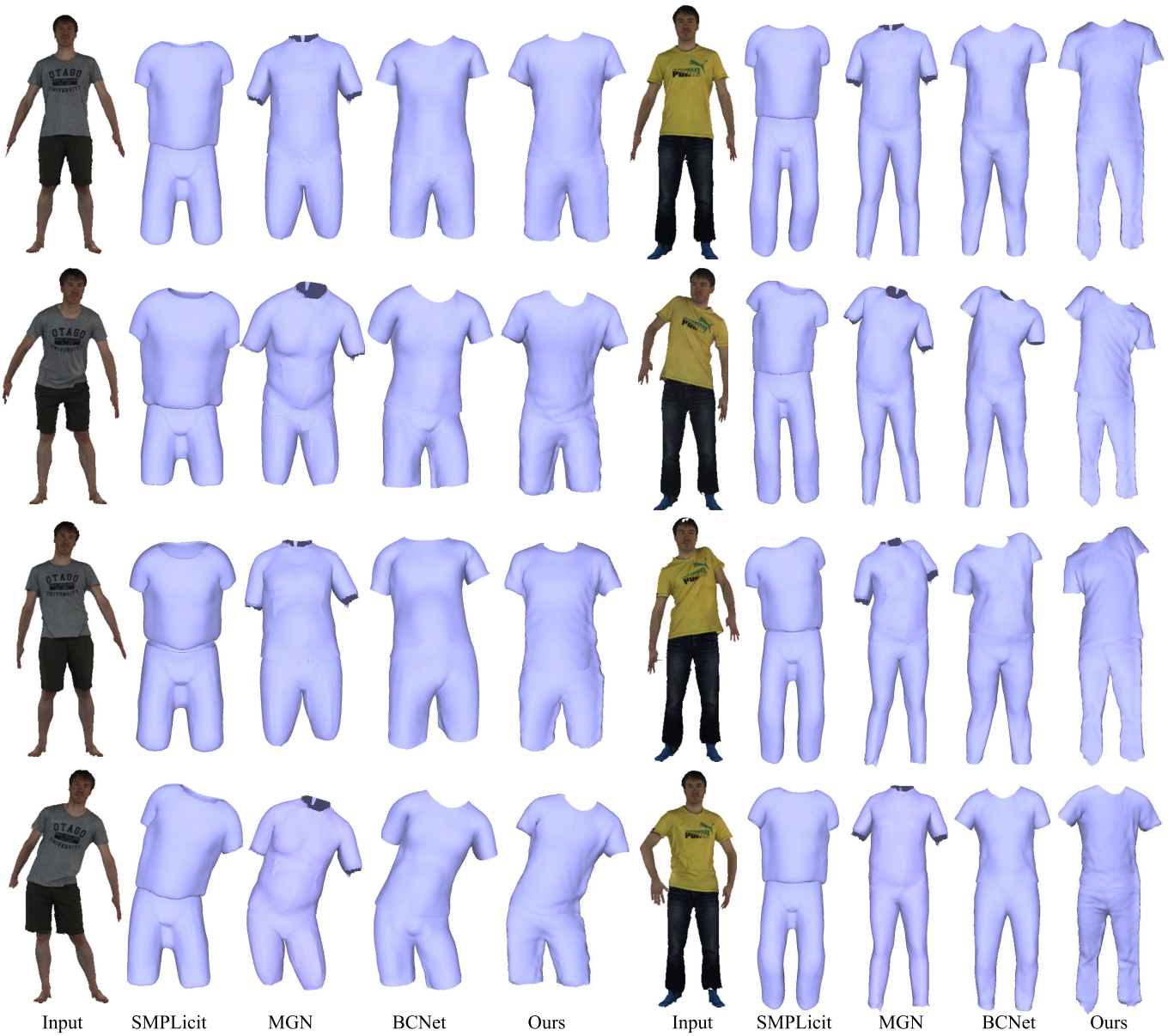

Fig. 4. Reconstructed garments by SMPLicit [25], MGN [22], BCNet [23] and our method on CAPE dataset [74]. The inputs are four frames of a motion sequence.

during rotation, we only show the results of the first frame of the video. It can be seen that MGN [22] and SMPLicit [25] cannot get accurate clothing styles. While BCNet [23] can get visually reasonable shapes, it cannot produce geometric details that are consistent with the input, or even produces wrong details. On the contrary, our approach benefits from the weakly supervised framework and reconstructs high-quality garments which faithfully reflect the input appearances. The elegant design of the multi-hypothesis displacement module also enables back surfaces with reasonable details to be generated, given the input of the front view. To further demonstrate the effect of our model, we also compare our method with a video-based method REC-MV [61]. In the current REC-MV source code, there is an absence of data preprocessing code, *e.g.*, estimating the feature lines of the clothes from the input, which is based on their previous work called Deep Fashion3D [24] and is currently not accessible. Therefore, we can only make a qualitative comparison on the People-Snapshot dataset, because the preprocessing data of the People-Snapshot dataset is released. In Fig. 6, since the person maintains a rough A-pose during rotation, we only show the results of the first frame of the video. It can be seen that our model can not only reconstruct the garment geometry consistent with the image, but also keep the garment stable. Compared to our method, the training time of REC-MV is around 18 hours with an RTX 3090 GPU, while we train our model with a TITAN X GPU in half an hour. Besides, the results of REC-MV could not be animated. More dynamic results can be found in supplementary video.

2) Quantitative Comparison: We test our method and the state-of-the-art methods with the rendered images from CAPE [74] dataset. Note that we use all the subjects with raw

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

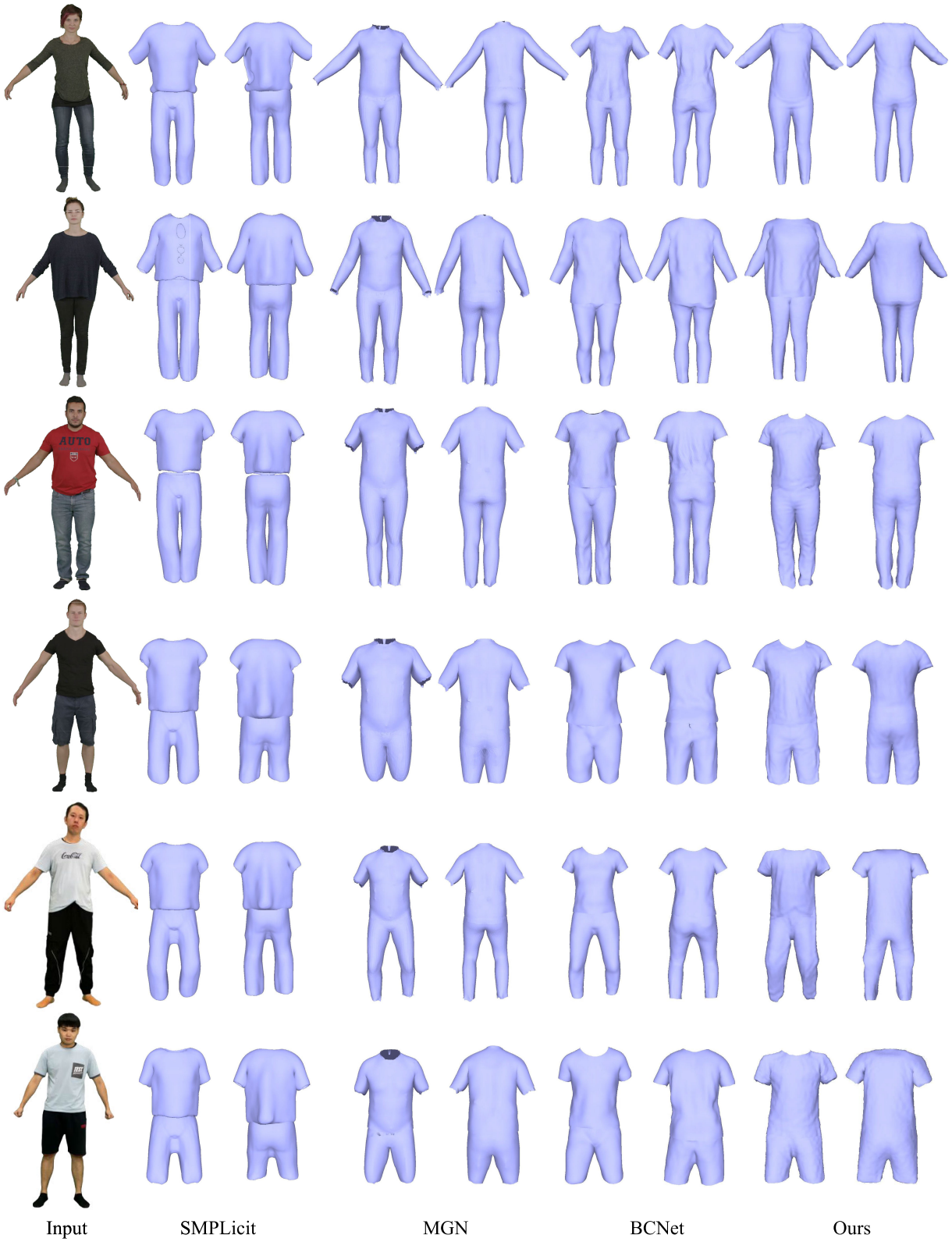

Fig. 5. Reconstructed garments by SMPLicit [25], MGN [22], BCNet [23] and our method on People-Snapshot [18] dataset (top four rows) and our captured data (bottom two rows).

scans ('00032-shortshort-hips', '00096-shortshort-tilt-twist-left', '00159-shortlong-pose-model', '03223-shortlong-hips') from CAPE [74] dataset, and for brevity, only the ID of the subject is kept in the table in the rest of this section. We first align the garment meshes generated by different methods to the ground truth meshes across all frames for a video and then compute the final average Chamfer distance between the reconstructed garments and the ground truth meshes for

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

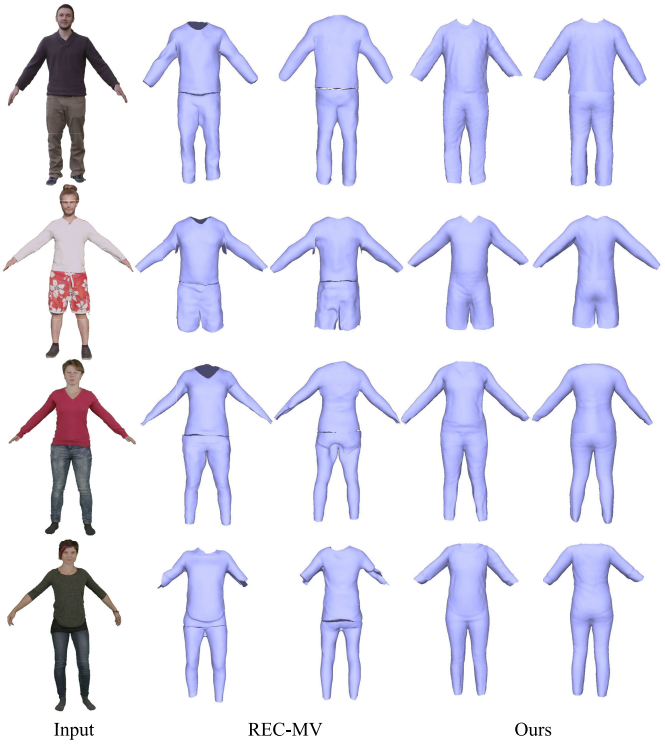

Fig. 6. Reconstructed garments by REC-MV [61] and our method on People-Snapshot [18] dataset.

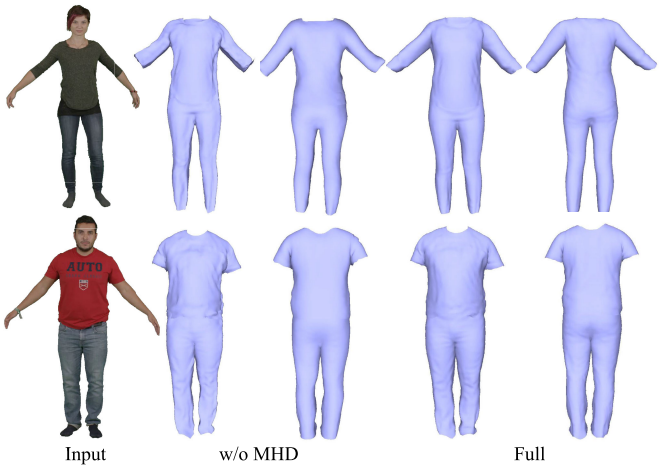

Fig. 7. Qualitative results of multi-hypothesis displacement module ablation study.

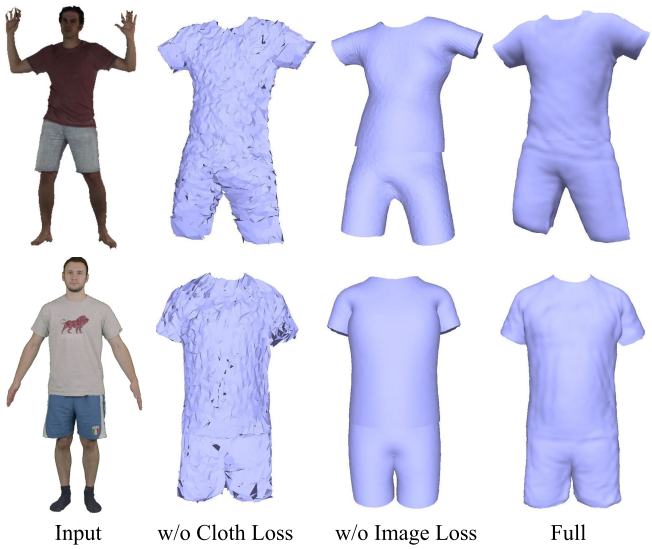

Fig. 8. Qualitative results of loss function ablation study.

Fig. 9. Qualitative results of garment template ablation study.

accuracy measurement. To evaluate the temporal consistency of the reconstructed meshes, we measure the consistency of corresponding vertices (CCV), which is the root mean squared error of the corresponding vertices' distances in adjacent frames. As shown in Table I, our method outperforms other methods in reconstruction accuracy, which indicates more realistic reconstruction results from a single RGB camera.

C. Ablation Study

1) Multi-Hypothesis Displacement: To validate the effect of the multi-hypothesis displacement module, we compare the performances of using different numbers of hypotheses. Specifically, given the high-level embedding of the pose, we define different numbers of learnable matrices to get deformations and encourage gradient propagation through residual connections. Then, we connect these hypothetical deformations as the input of an MLP to output the final deformation. Table II gives the quantitative results on CAPE dataset [74]. We calculate the average Chamfer distance between the aligned reconstructed garments and the ground truth meshes across all frames for a video and consistency of corresponding vertices (CCV) between adjacent frames. As shown in Table II, different numbers of hypotheses achieve similar accuracies, and all have higher accuracies than w/o MHD. In the rest of this section, we utilize MHD3 as our full model. Some visual results are shown in Fig. 7. It can be

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

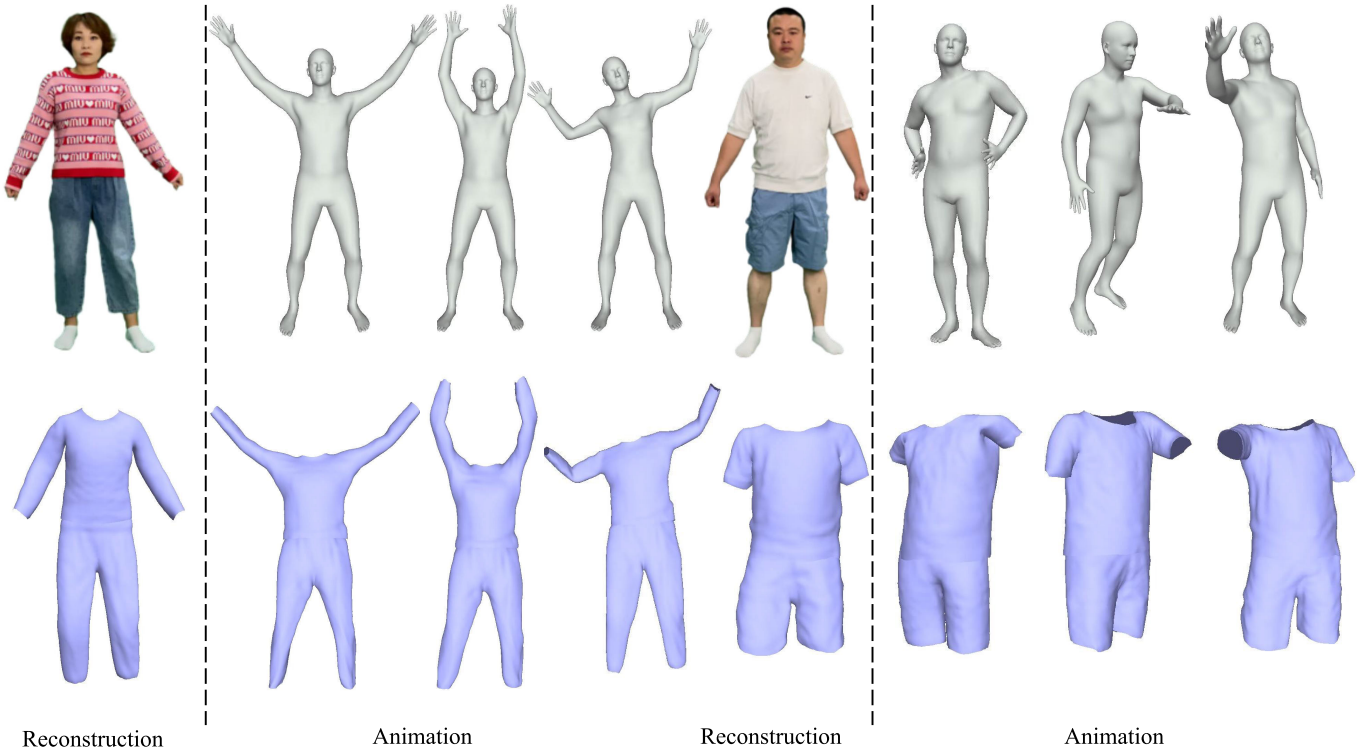

Fig. 10. Garment animation results.

TABLE I
QUANTITATIVE COMPARISON ON CAPE DATASET

Method	00032		00096		00159		03223	
	CD↓	CCV↓	CD↓	CCV↓	CD↓	CCV↓	CD↓	CCV↓
SMPLicit	1.611	-	1.866	-	1.811	-	1.599	-
MGN	1.328	2.927	1.850	2.055	1.345	2.959	1.452	2.983
BCNet	1.591	3.877	1.240	4.212	1.270	2.305	1.477	2.350
Ours	**1.098**	**1.819**	**1.049**	**1.217**	**1.087**	**0.961**	**1.072**	**0.713**

TABLE II
QUANTITATIVE EVALUATION FOR MULTI-HYPOTHESIS DISPLACEMENT MODULE ABLATION STUDY (CM)

Method	00032		00096		00159		03223	
	CD↓	CCV↓	CD↓	CCV↓	CD↓	CCV↓	CD↓	CCV↓
w/o MHD	1.167	1.835	1.130	1.204	1.198	1.076	1.159	0.743
MHD2	1.105	1.828	1.051	1.203	1.090	0.968	1.079	0.719
MHD3	**1.098**	**1.819**	**1.049**	1.217	1.087	0.961	**1.072**	**0.713**
MHD4	1.110	1.829	1.059	**1.193**	1.087	0.972	1.086	0.718
MHD5	1.106	1.820	1.057	1.194	**1.083**	**0.960**	1.089	0.720
MHD6	1.105	1.828	1.056	1.206	1.089	0.961	1.085	0.718

seen that our full model addresses the problems faced by w/o MHD, such as messy details and over-smooth back surfaces. At the same time, it also proves the effectiveness of our multi-hypothesis module, which can learn the dynamic deformations of clothes well from monocular video.

2) Loss Function: We study the effects of different loss functions on garment reconstruction. Our method is compared with two variants: one supervised without clothing loss function (w/o Cloth Loss), and the other supervised without image loss function (w/o Image Loss). In the same way as before, we calculate the average Chamfer distance between the aligned reconstructed garments and the ground truth meshes across all frames for a video ywand consistency of corresponding vertices(CCV) between adjacent frames. Table III gives the quantitative results in terms of the Chamfer distance and consistency of corresponding vertices (CCV). Our full model achieves the best performance, which verifies the importance of adopting both the image loss and the clothing loss. As shown in Fig. 8, the variant without the clothing loss generates messy meshes, while the variant without the image loss generates smooth meshes. In contrast, our full model can not only reconstruct the garment geometry consistent with the image, but also keep the garment stable.

TABLE III
QUANTITATIVE EVALUATION FOR LOSS FUNCTION ABLATION STUDY (CM)

Method	00032		00096		00159		03223	
	CD↓	CCV↓	CD↓	CCV↓	CD↓	CCV↓	CD↓	CCV↓
w/o Cloth Loss	1.397	1.844	1.552	1.243	1.677	**0.924**	1.580	0.725
w/o Image Loss	1.172	**1.808**	1.128	1.275	1.174	0.967	1.158	0.781
Full	**1.098**	1.819	**1.049**	**1.217**	**1.087**	0.961	**1.072**	**0.713**

3) Garment Template: We study the effects of different parametric garment templates on garment reconstruction. We compare our method with a variant template generated by MGN (Ours-MGN). In the same way as before, we calculate the average Chamfer distance between the aligned reconstructed garments and the ground truth meshes across all frames for a video and consistency of corresponding vertices (CCV) between adjacent frames. Table IV gives the quantitative results in terms of the Chamfer distance and consistency of corresponding vertices (CCV). Our full model achieves slightly better performance than Ours-MGN. Both

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

Fig. 11. Neural texture generation by our method on Cape dataset (left three columns) and People-Snapshot dataset (right three columns).

Ours-MGN and our full model outperform other state-of-the-art (SOTA) methods. As shown in Fig. 9, Ours-MGN also reconstructs the garment geometry consistent with the image, but at the neckline and cuffs, there are a lot of messy triangle faces.

TABLE IV
QUANTITATIVE EVALUATION FOR GARMENT TEMPLATE ABLATION STUDY (CM)

Method	00032		00096		00159		03223	
	CD↓	CCV↓	CD↓	CCV↓	CD↓	CCV↓	CD↓	CCV↓
Ours-MGN	1.246	1.896	1.078	**0.968**	1.135	1.039	1.201	0.979
Full	**1.098**	**1.819**	**1.049**	1.217	**1.087**	**0.961**	**1.072**	**0.713**

D. Garment Animation

We utilize a parametric body model of SMPL [27] which makes the garment deformation more controllable, in order to handle dynamic garment reconstruction from monocular videos. Thanks to the design of the learnable garment deformation network, our method can generate reasonable deformations for unseen poses. Specifically, we train the model using our captured videos and test it with random unseen pose sequences. Table V gives the quantitative results in terms of the

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

TABLE V
QUANTITATIVE EVALUATION FOR GARMENT ANIMATION (CM)

Method	00032		00096		00159		03223	
	CD↓	CCV↓	CD↓	CCV↓	CD↓	CCV↓	CD↓	CCV↓
reconstruction	**1.098**	**1.819**	**1.049**	**1.217**	**1.087**	**0.961**	**1.072**	**0.713**
animation	1.103	2.543	1.081	2.065	1.097	1.076	1.112	0.743

Fig. 12. Reconstructed loose garment by MGN [22], BCNet [23] and our method on our captured data.

Chamfer distance, as well as the consistency of corresponding vertices (CCV) between adjacent frames. As shown in the Table V, our method achieves similar accuracy on unseen poses. Figure 10 shows that our method can still produce garments with well-preserved personal identity and clothing details of the subjects under various novel poses, which enables dynamic garment animation. More dynamic results can be found in supplementary video.

E. Texture Generation

Figure 11 shows some qualitative results by our neural texture generation method. As shown in the figure, our method can not only obtain high-quality garment geometry, but also produce high-fidelity textures consistent with the image. More dynamic results can be found in the supplementary video.

F. Discussion and Limitations

Although we have achieved high-quality animatable dynamic garment reconstruction from a single RGB camera, there are still some cases that we cannot solve well:

1) Loose Clothing: The results of cases with loose clothes may not be good, due to less relevance between body and clothing. Figure 12 shows some comparison results. Our method can obtain visually reasonable clothing shapes, but cannot recover folded structures consistent with the image. In further work, we will design a temporal fusion module that uses information from adjacent frames to improve the representation of the framework and generate higher quality animatable dynamic garments.

2) Collars: While our method can reconstruct garment meshes with high-quality surface details from a monocular video, it fails to reconstruct collars due to the lack of supervision of the collars. Fig. 14 gives some examples of such cases. We will explore a post-processing to extend our method to address this.

3) Extreme Poses: Although our method can generate reasonable deformations for unseen poses, it may produce incorrect results for extreme poses. Fig. 14 gives some examples of extreme poses cases. This could be solved by adding garment priors and training with more poses in the future work.

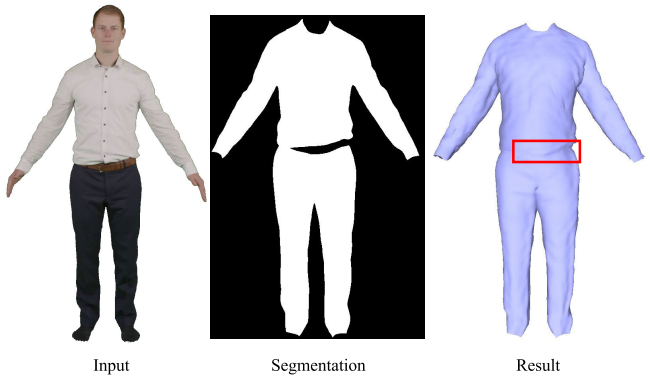

Fig. 13. An example of imprecisely reconstructed clothing due to wrong segmentation result.

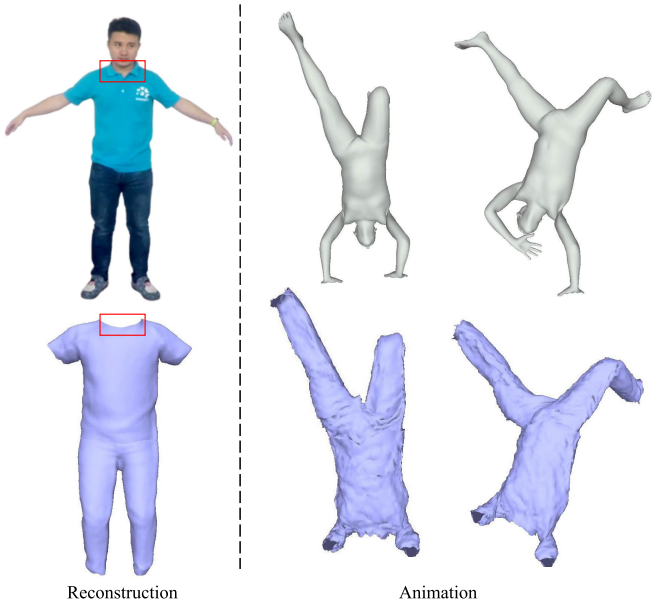

Fig. 14. Examples of failure cases for collars and extreme poses.

4) Segmentation and Normal Estimation: By applying weakly supervised training, we eliminate the need to simulate or scan hundreds or even thousands of sequences. Instead, our method uses predicted clothing segmentation masks and normal maps as the 2D supervision during training. The errors of segmentation and normal estimation could have negative effects on the training process and lead to imprecise reconstruction. Fig. 13 gives some examples of wrong clothing segmentation.

VI. CONCLUSION

In this paper, we aim to solve a meaningful but challenging problem: reconstructing high-quality animatable dynamic garments from monocular videos. We propose a weakly supervised framework to eliminate the need to simulate or scan hundreds or even thousands of sequences. To the best of our knowledge, no other work meets the expectations of digitizing high-quality garments that can be adjusted to

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

various unseen poses. In particular, we propose a learnable garment deformation network that formulates the garment reconstruction task as a pose-driven deformation problem. This design enables our model to generate reasonable deformations for various unseen poses. To alleviate the ambiguity brought by estimating 3D garments from monocular videos, we design a multi-hypothesis deformation module that learns spatial representations of multiple plausible deformation hypotheses. In this way, we can alleviate the ambiguity brought by estimating 3D garments based on monocular videos. The prior works [22], [23], [25] focus on the geometry of the clothes and do not attempt to recover the garment textures, which limits their application scenarios. Therefore, we design a neural texture network to generate high-fidelity textures consistent with the image. Experimental results on several public datasets demonstrate that our method can reconstruct high-quality dynamic garments with coherent surface details, which can be easily animated under unseen poses.

REFERENCES

[1] Y. Fu, R. Li, T. S. Huang, and M. Danielsen, "Real-time multimodal human–avatar interaction," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 18, no. 4, pp. 467–477, Apr. 2008.

[2] J. Yang, X. Guo, K. Li, M. Wang, Y.-K. Lai, and F. Wu, "Spatio-temporal reconstruction for 3D motion recovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1583–1596, Jun. 2020.

[3] N. Jojic, J. Gu, T. S. Shen, and T. S. Huang, "Computer modeling, analysis, and synthesis of dressed humans," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 2, pp. 378–388, Mar. 1999.

[4] O. Sarakatsanos et al., "A VR application for the virtual fitting of fashion garments on avatars," in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, Oct. 2021, pp. 40–45.

[5] A. Genay, A. Lécuyer, and M. Hachet, "Being an avatar 'for real': A survey on virtual embodiment in augmented reality," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 12, pp. 5071–5090, Dec. 2022.

[6] S. Pujades et al., "The virtual caliper: Rapid creation of metrically accurate avatars from 3D measurements," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 5, pp. 1887–1897, May 2019.

[7] Y. Xu, S. Yang, W. Sun, L. Tan, K. Li, and H. Zhou, "3D virtual garment modeling from RGB images," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2019, pp. 37–45.

[8] X. Li et al., "Learning to infer inner-body under clothing from monocular video," *IEEE Trans. Vis. Comput. Graphics*, early access, Aug. 29, 2022, doi: 10.1109/TVCG.2022.3202240.

[9] Y. Li, M. Habermann, B. Thomaszewski, S. Coros, T. Beeler, and C. Theobalt, "Deep physics-aware inference of cloth deformation for monocular human performance capture," in *Proc. Int. Conf. 3D Vis. (3DV)*, Dec. 2021, pp. 373–384.

[10] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3D scan sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4191–4200.

[11] X. Chen, A. Pang, W. Yang, P. Wang, L. Xu, and J. Yu, "TightCap: 3D human shape capture with clothing tightness field," *ACM Trans. Graph.*, vol. 41, no. 1, pp. 1–17, Feb. 2022.

[12] Z. Lahner, D. Cremers, and T. Tung, "DeepWrinkles: Accurate and realistic clothing modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 667–684.

[13] K. Li et al., "SPA: Sparse photorealistic animation using a single RGB-D camera," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 771–783, Apr. 2017.

[14] D. S. Alexiadis et al., "An integrated platform for live 3D human reconstruction and motion capturing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 798–813, Apr. 2017.

[15] P. Cong, Z. Xiong, Y. Zhang, S. Zhao, and F. Wu, "Accurate dynamic 3D sensing with Fourier-assisted phase shifting," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 3, pp. 396–408, Apr. 2015.

[16] J. Peng, Z. Xiong, Y. Zhang, D. Liu, and F. Wu, "LF-fusion: Dense and accurate 3D reconstruction from light field images," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[17] H. Zhu, Y. Liu, J. Fan, Q. Dai, and X. Cao, "Video-based outdoor human reconstruction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 760–770, Apr. 2017.

[18] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3D people models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8387–8397.

[19] S. Saito, Z. Huang, R. Natsume, S. Morishima, H. Li, and A. Kanazawa, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2304–2314.

[20] Q. Feng, Y. Liu, Y.-K. Lai, J. Yang, and K. Li, "FOF: Learning Fourier occupancy field for monocular real-time human reconstruction," 2022, *arXiv:2206.02194*.

[21] H. Zhao et al., "High-fidelity human avatars from a single RGB camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15904–15913.

[22] B. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-garment net: Learning to dress 3D people from images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5420–5430.

[23] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, and H. Bao, "BCNet: Learning body and cloth shape from a single image," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 18–35.

[24] H. Zhu et al., "Deep Fashion3D: A dataset and benchmark for 3D garment reconstruction from single images," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 512–530.

[25] E. Corona, A. Pumarola, G. Alenyà, G. Pons-Moll, and F. Moreno-Noguer, "SMPLicit: Topology-aware generative model for clothed people," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11875–11885.

[26] H. Zhu, L. Qiu, Y. Qiu, and X. Han, "Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3835–3844.

[27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, 2015.

[28] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Detailed human avatars from monocular video," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 98–109.

[29] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1175–1186.

[30] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2Shape: Detailed full human body geometry from a single image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2293–2303.

[31] V. Lazova, E. Insafutdinov, and G. Pons-Moll, "360-degree textures of people in clothing from a single image," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 643–653.

[32] J. Hu, H. Zhang, Y. Wang, M. Ren, and Z. Sun, "Personalized graph generation for monocular 3D human pose and shape estimation," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 31, 2023, doi: 10.1109/TCSVT.2023.3310525.

[33] L. Wang, X. Liu, X. Ma, J. Wu, J. Cheng, and M. Zhou, "A progressive quadric graph convolutional network for 3D human mesh recovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 104–117, Jan. 2023.

[34] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, "Detailed human shape estimation from a single image by hierarchical mesh deformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4491–4500.

[35] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "DeepHuman: 3D human reconstruction from a single image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7739–7749.

[36] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 165–174.

[37] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 84–93.

[38] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung, "ARCH: Animatable reconstruction of clothed humans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3093–3102.

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.

[39] T. He, J. Collomosse, H. Jin, and S. Soatto, "Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 9276–9287.

[40] T. He, Y. Xu, S. Saito, S. Soatto, and T. Tung, "ARCH++: Animation-ready clothed human reconstruction revisited," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11046–11056.

[41] Z. Zheng, T. Yu, Y. Liu, and Q. Dai, "PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3170–3184, Jun. 2022.

[42] Y. Hong, J. Zhang, B. Jiang, Y. Guo, L. Liu, and H. Bao, "StereoPIFu: Depth aware clothed human digitization via stereo vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 535–545.

[43] Y. Cao, G. Chen, K. Han, W. Yang, and K. K. Wong, "JIFF: Jointly-aligned implicit face function for high quality single view clothed human reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2729–2739.

[44] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 27171–27183.

[45] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 405–421.

[46] B. Jiang, Y. Hong, H. Bao, and J. Zhang, "SelfRecon: Self reconstruction your digital avatar from monocular video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5605–5615.

[47] C. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman, "HumanNeRF: Free-viewpoint rendering of moving people from monocular video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16210–16220.

[48] S.-Y. Su, F. Yu, M. Zollhöfer, and H. Rhodin, "A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12278–12291.

[49] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan, "NeuMan: Neural human radiance field from a single video," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 402–418.

[50] Z. Li, Z. Zheng, H. Zhang, C. Ji, and Y. Liu, "AvatarCap: Animatable avatar conditioned monocular human volumetric capture," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 322–341.

[51] J. Chen et al., "Animatable neural radiance fields from monocular RGB videos," 2021, *arXiv:2106.13629*.

[52] G. Moon, H. Nam, T. Shiratori, and K. M. Lee, "3D clothed human reconstruction in the wild," 2022, *arXiv:2207.10053*.

[53] F. Zhao, W. Wang, S. Liao, and L. Shao, "Learning anchored unsigned distance functions with gradient direction alignment for single-view garment reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12674–12683.

[54] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–15, Aug. 2017.

[55] D. Xiang et al., "Dressing avatars: Deep photorealistic appearance for physically simulated clothing," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–15, Dec. 2022.

[56] G. Tiwari, B. L. Bhatnagar, T. Tung, and G. Pons-Moll, "Sizer: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–18.

[57] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4460–4470.

[58] O. Halimi et al., "Garment avatars: Realistic cloth driving using pattern registration," 2022, *arXiv:2206.03373*.

[59] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt, "DeepCap: Monocular human performance capture using weak supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5052–5063.

[60] Y. Feng, J. Yang, M. Pollefeys, M. J. Black, and T. Bolkart, "Capturing and animation of body and clothing from monocular video," 2022, *arXiv:2210.01868*.

[61] L. Qiu, G. Chen, J. Zhou, M. Xu, J. Wang, and X. Han, "REC-MV: Reconstructing 3D dynamic cloth from monocular videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4637–4646.

[62] G. Pavlakos et al., "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10975–10985.

[63] Z. Yao, K. Li, Y. Fu, H. Hu, and B. Shi, "GPS-Net: Graph-based photometric stereo network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 10306–10316.

[64] Y. Ju, B. Shi, M. Jian, L. Qi, J. Dong, and K.-M. Lam, "NormAttention-PSN: A high-frequency region enhanced photometric stereo network with normalized attention," *Int. J. Comput. Vis.*, vol. 130, no. 12, pp. 3014–3034, Dec. 2022.

[65] G. Chen, K. Han, B. Shi, Y. Matsushita, and K. K. Wong, "Deep photometric stereo for non-Lambertian surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 129–142, Jan. 2022.

[66] Y. Liu et al., "A deep-shallow and global–local multi-feature fusion network for photometric stereo," *Image Vis. Comput.*, vol. 118, Feb. 2022, Art. no. 104368.

[67] H. Zhang et al., "PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11446–11456.

[68] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 770–785.

[69] J. Zhang, L. Gu, Y.-K. Lai, X. Wang, and K. Li, "Towards grouping in large scenes with occlusion-aware spatio-temporal transformers," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Oct. 16, 2023, doi: 10.1109/TCSVT.2023.3324868.

[70] P. Henderson and V. Ferrari, "Learning single-image 3D reconstruction by generative modelling of shape, pose and shading," *Int. J. Comput. Vis.*, vol. 128, no. 4, pp. 835–854, Apr. 2020.

[71] H. Bertiche, M. Madadi, and S. Escalera, "PBNS: Physically based neural simulator for unsupervised garment pose space deformation," 2020, *arXiv:2012.11310*.

[72] I. Santesteban, M. A. Otaduy, and D. Casas, "SNUG: Self-supervised neural dynamic garments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8140–8150.

[73] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[74] Q. Ma et al., "Learning to dress 3D people in generative clothing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6469–6478.

[75] K.-A. Aliev, A. Sevastopolsky, M. Kolos, D. Ulyanov, and V. Lempitsky, "Neural point-based graphics," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 696–712.

[76] S. Prokudin, M. J. Black, and J. Romero, "SMPLpix: Neural avatars from 3D human models," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1810–1819.

[77] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5904–5913.

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on November 28,2024 at 06:56:24 UTC from IEEE Xplore. Restrictions apply.