Clothed Human Performance Capture with a Double-layer Neural Radiance Fields

Kangkan Wang*[1,2], Guofeng Zhang[3], Suxu Cong[1] , Jian Yang[1,2]
[1]Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education,
[2]Jiangsu Key Lab of Image and Video Understanding for Social Security,
School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[3]State Key Laboratory of CAD&CG, Zhejiang University, China

Abstract

This paper addresses the challenge of capturing performance for the clothed humans from sparse-view or monocular videos. Previous methods capture the performance of full humans with a personalized template or recover the garments from a single frame with static human poses. However, it is inconvenient to extract cloth semantics and capture clothing motion with one-piece template, while single frame-based methods may suffer from instable tracking across videos. To address these problems, we propose a novel method for human performance capture by tracking clothing and human body motion separately with a double-layer neural radiance fields (NeRFs). Specifically, we propose a double-layer NeRFs for the body and garments, and track the densely deforming template of the clothing and body by jointly optimizing the deformation fields and the canonical double-layer NeRFs. In the optimization, we introduce a physics-aware cloth simulation network which can help generate physically plausible cloth dynamics and body-cloth interactions. Compared with existing methods, our method is fully differentiable and can capture both the body and clothing motion robustly from dynamic videos. Also, our method represents the clothing with an independent NeRFs, allowing us to model implicit fields of general clothes feasibly. The experimental evaluations validate its effectiveness on real multi-view or monocular videos.

1. Introduction

Performance capture for clothed humans is one of the essential problems in the metaverse, and it not only captures the inner human body motion but also recovers the outer clothing motion which has many promising applications such as virtual try-on, video editing, and telepresence. From sparse-view or monocular videos of a moving human in general clothes, its goal is to recover the dynamic 3D shape sequence of the human body and clothing simultaneously that are consistent with the observed frames in both human shape and motion. This is a very challenging problem since the dynamic human could be with arbitrary motions and with complex non-rigid cloth deformations, and the clothing in motion is difficult to maintain physically plausible interactions with the body.

Previous systems [4, 21, 38, 39] reconstruct 3D clothed humans by using depth sensors or fitting a personalized template [9, 10, 37] to the image observations (e.g., body joints and silhouettes). Only recovering one-piece geometry which unifies the human body and clothes, these systems fail to track the motion of the clothing and achieve clothing editing on 3D humans, which are the prerequisites in many VR/AR applications like virtual dressing. On the contrary, cloth can be extracted and tracked from depth scans [26, 40] accurately by fitting pre-built cloth templates to the scans which have limited applications when 3D data are unavailable. Existing garment estimation methods [3, 12, 43, 44] from color images require the person facing the camera and in static poses. When the human is moving and the clothing is deforming, these methods may recover the 3D garments unreliably. Recent methods [16, 31] track body and clothing motion simultaneously from videos, but they need to re-build cloth template for a new performer and the running efficiency is very low due to online cloth simulation or computationally-exhaustive optimization, which prohibits them from being widely deployed for daily applications.

Recent works [20, 34, 41] adopt dynamic human NeRFs to capture human motion and obtain impressive tracking results. By capturing the temporally-varying appearance in the videos, dynamic NeRFs [34] can provide dense photometric constraints to track the deforming geometry of the performer. However, they represent the human with a single NeRFs without modeling cloth, and the clothing motion cannot be extracted. In this paper, we aim to track the cloth-

*Corresponding author: wangkangkan@njust.edu.cn

ing and body motion simultaneously with dynamic NeRFs. However, this problem is rather challenging due to two major questions we need to solve: how to represent dynamic clothing and human body with NeRFs, and how to capture clothing and human body motion with plausible body-cloth interactions based on the implicit representation.

In this paper, we propose a novel method for clothed human performance capture with a double-layer NeRFs. Specifically, a double-layer NeRFs is modeled for both the body and clothing in the template space, and transformed to the observation space with the corresponding deformation fields, and the rendered images are then synthesized by composing the two dynamic NeRFs. We first estimate the template shape in canonical frame and learn the geometry network supervised by the template geometry. In the rendering, we compose the double-layer NeRFs with the guidance of deformed body and clothing meshes. Then, by minimizing the difference between the rendered color and observed color, the deformation fields and the canonical NeRFs are optimized jointly. The deformation field is represented as the inverse deformation of the template mesh, thus the densely deforming geometry of the template can be recovered simultaneously. In addition, we adopt a physics-aware network learnt from simulation data between various cloth types and humans to constrain the dynamic clothing and preserve physically plausible body-cloth interactions, resulting in realistic cloth geometry tracking. Compared to previous methods, our method is fully differentiable and can recover the realistic motion of both the clothing and body from dynamic videos with arbitrary human poses and complex cloth deformations. The experimental qualitative and quantitative results on datasets of DynaCap [8] and DeepCap [10] prove that the proposed approach can robustly and accurately capture the motion for clothed humans. In summary, the primary contributions of our work include:

- We propose a double-layer NeRFs for dynamic humans in general clothing, allowing us to model implicit humans with a variety of clothes (e.g., loose dresses).
- To the best of our knowledge, we propose the first framework to capture clothing motion separately from the human body using the double-layer NeRFs, which provides dense appearance constraints on the geometry tracking and improves the robustness and accuracy.
- A differentiable physics-aware network is learnt for different common garments and used to preserve physically plausible cloth deformations in motion capture.

2. Related Work

Clothed human reconstruction. Depth-based methods [4, 21, 38, 39] reconstruct 3D humans in realtime by fusing the depth maps in a volumetric way. However, depth-based methods cannot synthesize photo-realistic videos and have constrained applications in daily life due to data capture limitations, e.g., high power consumption and sensitive to sunlight. Based on deep learning techniques, some approaches predict 3D humans from color images through implicit function [27, 28], coarse-to-fine mesh deformation [45], surface normal optimization [36, 42], or UV mapping [2]. Although these methods can reconstruct human details, the recovered shapes do not contain human semantics like human pose, shape and clothing. To recover the clothing, some methods [3, 19] add displacements on vertices of SMPL model [18] which only support tight clothing and recover coarse-level geometry. Template-based methods [9, 10, 37] recover 3D human motions by non-rigidly fitting a person-specific template mesh to 2D observed features (e.g., human joints or silhouettes). With a personalized template, these methods obtain impressive performance capture. However, the recovered meshes integrate the body and clothing in one piece, which hinders their applications like garment retargeting and virtual trying-on.

Clothing motion capture. Traditional methods [26, 40] capture the cloth motion by precisely fitting a cloth template to 3D data from depth sensors, which cannot generalize to the video-based setting and are limited to applications in indoor scenes. Deep learning-based approaches [3, 12, 43, 44] estimate the 3D garment shape from color images. However, these methods require the human facing the camera with a static pose and do not try temporally tracking of dynamic garments. Some works [23, 29] predict the cloth shape conditioned on the human pose and shape but the shapes are not consistent with the observed data. Recent works [16, 31] capture both human body and clothing motion from dynamic videos. MulayCap [31] optimizes a simulated garment on SMPL model with semantic cloth segmentation, but their traditional simulation and optimization are both computationally expensive. DeepCloth [16] incorporates a physics-based simulator to supervise the cloth deformations estimated with multi-view joints and silhouettes as DeepCap [10]. However, their simulation layer is not differentiable and needs to adjust the cloth material types online, which is a tedious and time consuming step. Moreover, with the same limitation as DeepCap, they cannot recover the detailed cloth deformation based on the sparse features.

Implicit Human Reconstruction. Clothed humans can be reconstructed through implicit representation-based methods such as voxel representation [33, 42], implicit function [27, 28], or neural radiance fields (NeRFs) [20, 24, 25]. Recent works combine explicit and implicit representations to produce high-quality avatar [11] or reconstruct high-fidelity garment mesh [44], which show the advantage of implicit representation on recovering geometry details. However, these methods are limited to self-rotating human motion or a single color image. Neuralbody [25] and AniNeRF [24] extend NeRFs for a dynamic human with defor-

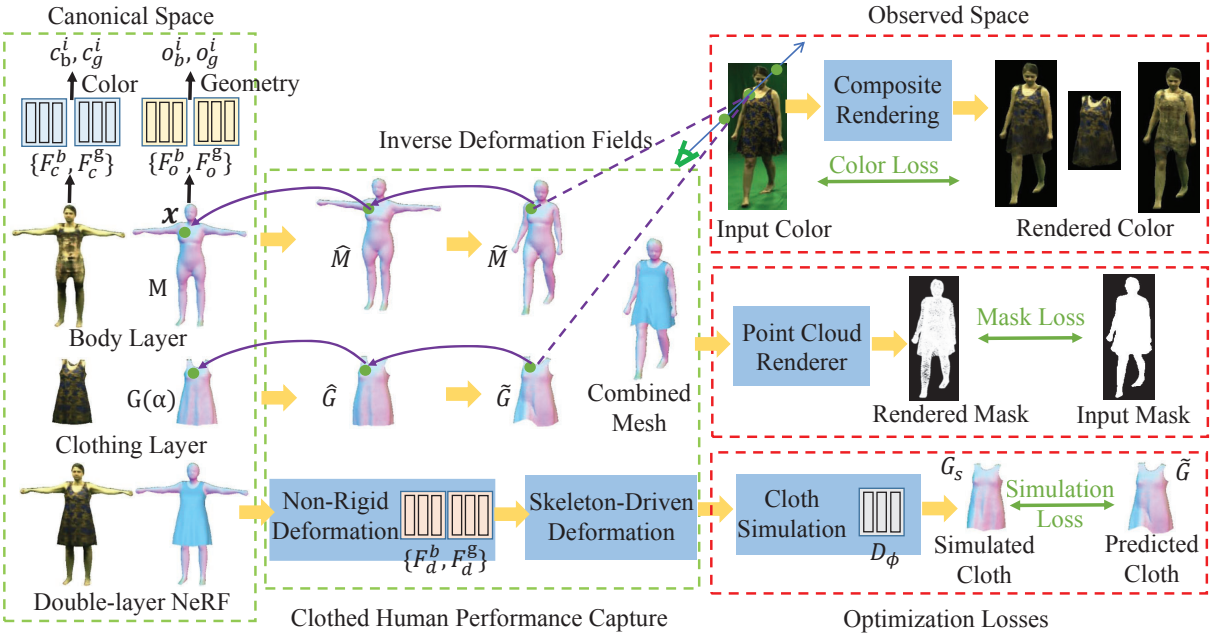


Figure 1. Overview of the proposed approach. Our approach can capture both the clothing and human body motion, and achieve novel-view synthesis from sparse-view or monocular videos. We learn a double-layer neural radiance fields (NeRFs) for the clothing and human body in canonical frame, and jointly optimize the canonical double-layer NeRFs and the human deformation constrained with a physics-aware simulation network. Please refer to Sec. 3 for details of our method.

mation fields based on the linear blend skinning (LBS) [15]. Some human NeRFs methods [17, 41] model nonrigid human deformations with a point displacement on top of LBS. However, without using any clothing motion prior, these methods cannot effectively recover the motion of humans in loose clothes, and the geometry of clothes is not reconstructed reliably under nonrigid cloth deformations. The recent work, NerfCap [34] incorporates the embedded graph prior into the human NeRFs and represents nonlinear surface deformations for loose clothes successfully. However, their method does not model the clothing separately and thus fails to track the clothing motion.

3. Proposed Approach

Given sparse-view or monocular videos of a clothed human, we track the motion of both the human body and clothing and synthesize novel-view videos of the human with a novel performance capture method based on a double-layer neural radiance fields (NeRFs). Figure 1 shows an overview of the proposed method. The double-layer NeRFs is learnt for the human body and clothing in canonical frame (Sec. 3.2). Then, the canonical NeRFs is dynamically transformed into the observed frame with inverse deformation fields of the deformed template for clothing and body respectively. From the dynamic double-layer NeRFs, we render the images using a composite rendering. By matching the synthesized images with the video frames and constraining the cloth deformations with a physics-aware simulation network, we estimate the double-layer NeRFs and the human deformation jointly (Sec. 3.3).

3.1. Canonical Human Shape

We decompose the clothed human into undressed body and garments (like T-shirt and pants), and represent them with SMPL model [18] and a parametric PCA model respectively. To include the human details (e.g., the head and hand), we obtain the canonical body model $\boldsymbol{M}$ by deforming the SMPL model to a personalized template with nonrigid deformation [32]. We design different styles (size and sleeve length, et al.) for each garment category, and simulate these garments on the template SMPL model using Marvelous Designer [1]. Then, we obtain a linear model through PCA on the simulated meshes,

$$\boldsymbol{G}(\boldsymbol{\alpha}) = \bar{\boldsymbol{G}} + \boldsymbol{B}_g\boldsymbol{\alpha}, \tag{1}$$

where $\bar{\boldsymbol{G}}$ is the mean garment shape, $\boldsymbol{B}_g$ denotes the linear basis matrix, and $\boldsymbol{\alpha}$ are parameters of the cloth shape $\boldsymbol{G}(\boldsymbol{\alpha})$.

3.2. Double-layer NeRFs for Dynamic Humans

Unlike a single NeRFs for clothed humans [17,24,25,34, 41], we represent the clothing with an independent NeRFs on the body, which forms a double-layer NeRFs. The implicit representation can be extended to multiple garments.

Canonical NeRF. We use an occupancy network [22] to represent canonical geometry since the occupancy can mod-

el geometry more precisely than density [20] as,

$$\boldsymbol{o}(\boldsymbol{x}) = \boldsymbol{F}_o(\gamma_{\boldsymbol{x}}(\boldsymbol{x})), \tag{2}$$

where $\boldsymbol{\gamma_x}$ is the positional encoding [20]. The color network $\boldsymbol{F}_c$ in canonical frame is formulated as,

$$\boldsymbol{c}_i(\boldsymbol{x}) = \boldsymbol{F}_c(\gamma_{\boldsymbol{x}}(\boldsymbol{x}), \boldsymbol{\varphi}_i), \tag{3}$$

where $\boldsymbol{\varphi}_i$ is an appearance latent code [24, 25] for frame i. We define occupancy network $\{\boldsymbol{F}_o^b, \boldsymbol{F}_o^g\}$, and color network $\{\boldsymbol{F}_c^b, \boldsymbol{\varphi}^b, \boldsymbol{F}_c^g, \boldsymbol{\varphi}^g\}$ for the body and clothing, respectively.

To facilitate the learning of occupancy network, we incorporate a signed distance fields (SDFs) network $\boldsymbol{S}$,

$$\boldsymbol{S} : \boldsymbol{p} \in \mathbb{R}^3 \rightarrow \boldsymbol{s} \in \mathbb{R}. \tag{4}$$

which predicts the SDF value $\boldsymbol{s}$ for point $\boldsymbol{p}$. We use the losses proposed in [30] to learn the SDFs of the shape,

$$\begin{aligned} L_{sdf} &= \sum_{\boldsymbol{p}\in\Phi} \mid \boldsymbol{S}(\boldsymbol{p}) \mid + \sum_{\boldsymbol{p}\in\Phi} (1 - \langle \nabla \boldsymbol{S}(\boldsymbol{p}), \bar{\boldsymbol{n}} \rangle) \\ &+ \sum_{\boldsymbol{p}\in\Omega} \mid \parallel \nabla \boldsymbol{S}(\boldsymbol{p}) \parallel_2 - 1 \mid + \sum_{\boldsymbol{p}\in\Omega\setminus\Phi} \exp(-\delta \cdot \mid \boldsymbol{S}(\boldsymbol{p}) \mid), \end{aligned} \tag{5}$$

where $\bar{\boldsymbol{n}}$ is the surface normal, $\nabla \boldsymbol{S}(\boldsymbol{p})$ indicates the gradient of the SDFs, Ω and Φ denotes the 3D space and shape surface respectively. The first and second terms are used to estimate valid SDFs and normals on the surfaces, respectively. The third term constrains the norm of SDF gradients to be 1, and the last term penalizes SDF values of non-surface points near 0 with $\delta \gg 1$. We define SDF network $\{\boldsymbol{S}^b, \boldsymbol{S}^g\}$ for both the body and clothing.

Supervised by the predictions of SDF network, the occupancy network is learnt with the following loss,

$$L_{occ} = \sum_{\boldsymbol{x}\in\Omega} L_C(\boldsymbol{F_o}(\gamma_{\boldsymbol{x}}(\boldsymbol{x})), \boldsymbol{o}), \tag{6}$$

where $\boldsymbol{o}$ is the occupancy indicated by SDFs (if $\boldsymbol{S}(\boldsymbol{x}) <= 0$, $\boldsymbol{o} = \boldsymbol{1}$; otherwise, $\boldsymbol{o} = \boldsymbol{0}$), and L_C denotes the cross-entropy loss. We adopt the compositional design of SDF and occupancy because it is better for volumetric learning with continuous occupancy fields [22] compared to discrete labels extracted from SDFs. The template is varying during learning geometry model, and the loss is formulated as,

$$L_{geo} = \mu_1 L_{sdf} + \mu_2 L_{occ}. \tag{7}$$

where μ_1 and μ_2 are the weight for each term. Through the optimization, the NeRF geometry is consistent with the estimated template so that densely deforming geometry of the template can be tracked by utilizing dense photometric constraints from NeRF color model.

Deformation Model. Following [34], we transform sampled points in observed space to canonical space with the inverse deformation of the nearest vertex on the deformed template. In this way, the human deformation is integrated into the NeRF optimization. The human deformation model is disentangled into the non-rigid surface deformation (i.e., embedded graph deformation as in [34]) and the articulated skeletal motion. For the clothing and body, we define the deformation network $\boldsymbol{F}_d^g$ with latent codes $\boldsymbol{\omega}^g$ and $\boldsymbol{F}_d^b$ with latent codes $\boldsymbol{\omega}^b$, respectively. Through nonrigid deformation, we obtain the clothing mesh $\hat{\boldsymbol{G}}$ and body mesh $\hat{\boldsymbol{M}}$. Then, the linear blend skinning [15] is applied to obtain the final cloth mesh $\tilde{\boldsymbol{G}}$ and body mesh $\tilde{\boldsymbol{M}}$ in the observation space. We recover the SMPL model with [13] for input video frames and also estimate a canonical body shape.

Composite Rendering. In [6], a blending weight is predicted for each sampled point to blend the static background and dynamic target in a scene. This blending method cannot separate the clothing and human body in the NeRFs effectively. We propose an efficient composite rendering strategy for the dynamic double-layer NeRF. Specifically, for a casted ray $\boldsymbol{r}$, we find the intersecting triangle with the deformed meshes. If the intersected triangle is from the body mesh, we set the mask $\boldsymbol{m}_r^b$ of all sampled points on $\boldsymbol{r}$ to 1; otherwise, $\boldsymbol{m}_r^b$ is set to 0. Similarly, if the intersected triangle is from the cloth mesh, we set the mask $\boldsymbol{m}_r^g$ to 1; otherwise, $\boldsymbol{m}_r^g$ is set to 0. The complete human is rendered as,

$$\begin{aligned} \boldsymbol{T}_r^i &= \prod_{j=1}^{i-1} (1 - \boldsymbol{o}_g^j (1 - \boldsymbol{m}_r^b))(1 - \boldsymbol{o}_b^j \boldsymbol{m}_r^b), \\ \hat{\boldsymbol{C}}_r &= \sum_{i=1}^{n} \boldsymbol{T}_r^i (\boldsymbol{o}_g^i (1 - \boldsymbol{m}_r^b) \boldsymbol{c}_g^i + \boldsymbol{o}_b^i \boldsymbol{m}_r^b \boldsymbol{c}_b^i), \end{aligned} \tag{8}$$

where $\{\boldsymbol{o}_g, \boldsymbol{o}_b\}$ and $\{\boldsymbol{c}_g, \boldsymbol{c}_b\}$ denote the point occupancy and color respectively, which are defined for both clothing and body. $\boldsymbol{T}_r^i$ is the accumulated transparency along the ray. We also render the clothing color $\hat{\boldsymbol{C}}_r^g$ and body color $\hat{\boldsymbol{C}}_r^b$.

3.3. Clothed Human Performance Capture with Double-layer NeRFs

We estimate the double-layer NeRFs and template deformation of the body and clothing by minimizing the following loss between the synthesized color and observed color:

$$\begin{aligned} L_{rgb} &= \sum_{\boldsymbol{r}\in\boldsymbol{R}} \|\hat{\boldsymbol{C}}(\boldsymbol{r}) - \boldsymbol{C}(\boldsymbol{r})\|_2^2 \\ &+ \sum_{\boldsymbol{r}\in\boldsymbol{R}} \|(\hat{\boldsymbol{C}}(\boldsymbol{r})^b - \boldsymbol{C}(\boldsymbol{r}))(1 - \boldsymbol{m}_r^g)\|_2^2 \\ &+ \sum_{\boldsymbol{r}\in\boldsymbol{R}} \|(\hat{\boldsymbol{C}}(\boldsymbol{r})^g - \boldsymbol{C}(\boldsymbol{r}))(1 - \boldsymbol{m}_r^b)\|_2^2, \end{aligned} \tag{9}$$

where $\boldsymbol{R}$ is the set of casted rays. While using the dense photometric constraint defined in Eq. 9, the model training

remains ill-posed and body-cloth collisions are prone to occur. To solve these problems, we design several regularization losses to constrain the joint optimization.

Physics-aware constraint. Cloth simulation is adopted in previous works [16,40] to obtain realistic performance capture. Different from their traditional simulation method, we introduce a differential network that learns a mapping from human and cloth parameters to physics-aware cloth shapes. We simulate the data in Marvelous Designer using the method of [23]. For each garment category, we use 25 garment styles $\boldsymbol{\gamma} \in \mathbb{R}^4$ as [23]. To account for the effects of different materials, we simulate 8 different materials in Marvelous Designer, and represent the material type $\boldsymbol{\tau}$ using one-hot labels. We train the simulation network $\boldsymbol{D}_\phi$ using a MLP to minimize the L1-loss between the simulated cloth meshes and the predicted ones $\boldsymbol{G}_s = \boldsymbol{D}_\phi(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau})$ with the input of human pose and shape, and garment and material styles. The physical simulation loss enforces the deformed cloth mesh $\tilde{\boldsymbol{G}}$ to be close to $\boldsymbol{G}_s$,

$$L_{sim} = \sum_{t \in \boldsymbol{G}} \rho(\|\tilde{\boldsymbol{G}}^t - \boldsymbol{G}_s^t\|_2^2), \tag{10}$$

where t denotes vertex index of $\boldsymbol{G}$, and ρ is the Geman-McClure robust function [5]. The simulation loss can penalize body-cloth collisions and preserve physically plausible deformations.

As-rigid-as-possible Loss. To make sure the template deforms smoothly, we utilize an as-rigid-as-possible loss [32] on the embedded graph deformation of the human body and clothing, $L_{arap} = L_{arap}(\hat{\boldsymbol{M}}) + L_{arap}(\hat{\boldsymbol{G}})$.

Mask Loss. We render the mask image $R(\boldsymbol{U})$ of the combined mesh $\boldsymbol{U} = (\boldsymbol{M}, \boldsymbol{G})$ by using a differentiable renderer [35] based on point cloud, and compute the mask loss with the input human mask image $\bar{R}$:

$$L_{mask}(\boldsymbol{U}) = \|R(\boldsymbol{U}) - \bar{R}\|_2^2. \tag{11}$$

We utilize the mask loss on both the deformed mesh and the simulated mesh as, $L_{mask} = L_{mask}(\tilde{\boldsymbol{M}}, \tilde{\boldsymbol{G}}) + L_{mask}(\tilde{\boldsymbol{M}}, \boldsymbol{G}_s)$.

Attachment Loss. During human motion, some cloth parts are always attached to the body (e.g., the dress strap is fixed on the shoulder). To ensure this constraint, we define an attachment loss:

$$L_{attach} = \sum_{i,j \in A} \rho(\|\tilde{\boldsymbol{G}}^i - \tilde{\boldsymbol{M}}^j\|_2^2), \tag{12}$$

where A is a set of selected vertices on the clothing mesh attached to corresponding points of the body mesh.

Interpenetration Loss. Interpenetrations may occur between the predicted garments and the body. We use the interpenetration term [7] to alleviate this problem:

$$L_{interp}(\boldsymbol{M}_1, \boldsymbol{M}_2) = \frac{1}{|C|} \sum_{i,j \in C} ReLU((\boldsymbol{M}_1^i - \boldsymbol{M}_2^j) \cdot \boldsymbol{N}_1^i), \tag{13}$$

where C is the set of corresponding points between the mesh $\boldsymbol{M}_1$ and mesh $\boldsymbol{M}_2$, $|C|$ denotes the number of corresponding points, and $\boldsymbol{N}_1$ is the normal of mesh $\boldsymbol{M}_1$. We use interpenetration loss in both canonical and posed space,

$$L_{interp} = L_{interp}(\hat{\boldsymbol{M}}, \hat{\boldsymbol{G}}) + L_{interp}(\tilde{\boldsymbol{M}}, \tilde{\boldsymbol{G}}). \tag{14}$$

3.3.1 Total Loss Function

We first learn the geometry network of both the clothing and human body in canonical frame. We use the first frame of multiple views or a video in self-rotating motion to estimate the canonical geometry by optimizing all the losses. Then, by fixing the clothing shape and the geometry network, we track the motion of each video frame and estimate the appearance of NeRFs using the following total loss:

$$\begin{aligned} L &= \lambda_1 L_{rgb} + \lambda_2 L_{arap} + \lambda_3 L_{sim} \\ &+ \lambda_4 L_{mask} + \lambda_5 L_{attach} + \lambda_6 L_{interp}, \end{aligned} \tag{15}$$

where $\{\lambda_1 ... \lambda_6\}$ are the balancing weights. For a video with N frames, the estimated parameter set X is:

$$X = \{\boldsymbol{F}_c^b, \boldsymbol{F}_c^g, \boldsymbol{F}_d^b, \boldsymbol{F}_d^g\} \cup \{\boldsymbol{\varphi}_i^b, \boldsymbol{\varphi}_i^g, \boldsymbol{\omega}_i^b, \boldsymbol{\omega}_i^g \mid i \in 1, ..., N\}, \tag{16}$$

which includes the weights of the appearance and deformation networks and the corresponding conditional variables for each frame in the video.

4. Experiments

4.1. Implementation details

Datasets and metrics. We design common garments (e.g., dress, T-shirt, skirt, pants, et al.) and simulate the data in the Marvelous Designer [1] under various human shapes and poses and cloth materials. We test our method on publicly datasets, DynaCap [8], and DeepCap [10]. We use four cameras around the human for training and the remaining cameras for test. Figure 2 shows some recovered results using our approach on different clothed humans. Our method is also tested on a monocular video in outdoor scene. Please refer to the supplementary video for more experimental results. The geometry accuracy is evaluated with the intersection over union (IoU) between the ground truth human masks and 2D projections of the recovered human shapes. Following [10, 34], the IoU (%) is computed on all views (AMVIoU), all views except the input view (RVIoU), and the input view (SVIoU). The novel-view synthesis quality is evaluated with two metrics, i.e., peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) as [24, 34].

Figure 2. The results of our approach on different humans in various garments. In each example, we show the ground truth image in a novel view, the reconstructed geometry (overlayed on the image), and the novel-view synthesis image. From left to right: the S4 from DeepCap dataset [10], "FranziRed" from DynaCap dataset [8], and the S1 from DeepCap dataset [10].

Architecture and experimental settings. We adopt the S-DF network from [30], and the same occupancy and color network as [34]. In the non-rigid deformation network, there are 8 layers with a skip connection at the fourth layer and the hidden size is 1024. The deformation and appearance latent codes both have dimensions of 128. We set both weight μ_1 and μ_2 to 1. We set the loss weight $\{\lambda_1, ..., \lambda_6\}$ to $\{1.0, 0.1, 0.02, 50, 0.1, 100.0\}$ empirically. The Adam optimizer [14] is used in the optimization, and the learning rate is $5e^{-4}$ at the start of training and exponentially decays to $5e^{-5}$ during the training. With a GPU of RTX 2080 Ti, our network training costs about 12 hours tested on a four-view video with 300 frames.

4.2. Comparison to state-of-the-art methods

We compare our method with two state-of-the-art approaches for human performance capture, DeepCap [10] and NerfCap [34]. We also compare with BCNet [12], TailorNet [23], and ICON [36]. For DeepCap and NerfCap, the experimental results are from [34]. For other compared methods, we implement them using their released codes. We do not compare with DeepCloth [16] since their codes are not released and DeepCap [10] outperforms them in geometry tracking as reported in [16]. We conduct the qualitative and quantitative experiments on S4 of DeepCap and

Method	AMVIoU	RVIoU	SVIoU
DeepCap [10]	86.17	85.75	86.90
NerfCap [34]	88.96	88.32	90.09
ICON [36]	-	-	**93.56**
BCNet [12]	82.33	81.67	82.94
TailorNet [23]	83.87	83.50	84.51
Our method	**89.58**	**88.75**	91.03

Evaluated on the subject S4 from DeepCap dataset [10].

Method	AMVIoU	RVIoU	SVIoU
DeepCap [10]	85.38	85.37	85.41
NerfCap [34]	88.08	88.07	88.12
ICON [36]	-	-	**95.21**
BCNet [12]	78.91	78.77	80.83
TailorNet [23]	78.15	78.00	78.21
Our method	**88.74**	**88.40**	89.68

Evaluated on the 'FranziRed" from DynaCap dataset [8].

Table 1. Comparison to the state-of-the-arts in surface reconstruction accuracy. Note that, the accuracy of ICON [36] is high since they use the human mask in estimating the geometry, and their dress shapes are not recovered accurately as shown in Fig. 3.

"FranziRed" from DynaCap. The test dataset are the same as [34] which include $300 - 400$ frames randomly sampled from the original dataset.

(a) (b) (c) (d) (e) (f) (g)

Figure 3. Qualitative comparison with the state-of-the-arts. (a) Ground truth image. (b) DeepCap [10]. (c) NerfCap [34]. (d) ICON [36]. (e) BCNet [12]. (f) TailorNet [23]. (g) Our method.

We show the qualitative comparisons in Figure 3 and list the quantitative results in Table 1. We only compute the error of the input view for ICON since they use their own camera system. Our method performs significantly better than the compared methods [12, 23, 36] without a cloth template. BCNet [12] also separates the modeling of the clothing and body, but as a supervised learning method, the performance may degrade dramatically because of data domain gap. When the human is moving, BCNet fails to recover the dresses accurately due to arbitrary cloth deformation. In contrast, the appearance loss of our double-layer NeRFs poses dense constraints on the clothing and body motion capture in a weakly-supervised manner. As a result, our method can achieve a much higher accuracy of geometry tracking even without the ground truth. The predicted garments of TailorNet [23] are not consistent with the real images since their model is conditioned on human poses and shapes. ICON obtains higher IoU accuracy by using the human mask in recovering the geometry. However, there are obvious artifacts on the recovered dresses of ICON [36]. Based on one-piece personalized template, it is hard for both DeepCap [10] and NerfCap [34] to extract the clothing motion, which is important for downstream applications in AR/VR. In contrast, our method tracks the motion for both the clothing and body. Without a pre-scanned cloth template, our method still obtains a high tracking accuracy thanks to the implicit representation.

Method	S4 of [10]		'FranziRed" of [8]	
	PSNR	SSIM	PSNR	SSIM
NerfCap [34]	**24.52**	**90.79**	**23.18**	**91.49**
Our method	23.06	90.28	21.70	90.54

Table 2. Comparison to NerfCap [34] in terms of novel-view synthesis on two datasets.

Method	AMVIoU	PSNR	SSIM
w/o simulation loss	89.03	22.66	89.35
w/o mask loss	88.75	23.21	89.49
w/o color loss	89.36	-	-
blending render [6]	88.75	23.64	90.51
1 camera view	88.49	21.34	87.12
2 camera views	89.23	21.89	88.38
6 camera views	**90.10**	**23.80**	**90.60**
ours(4 views)	89.58	23.06	90.28

Table 3. Ablation studies on S4 from DeepCap dataset [10].

We also compare the quality of free-viewpoint synthesis with NerfCap [34]. Figure 4 shows two compared examples and Table 2 reports the quantitative comparison. Without using any clothing template, our method still obtains close or comparable synthesis quality to NerfCap that uses a ground truth template. Our results are rendered by composing the NeRFs of the body and clothing, which is more complicated than a single NeRFs of NerfCap. For example, when the mutual motion occurs between the body and clothing (like cloth sliding), new body parts may appear, leading to the decrease of our synthesis quality.

(a) (b) (c) (d) (e) (f)

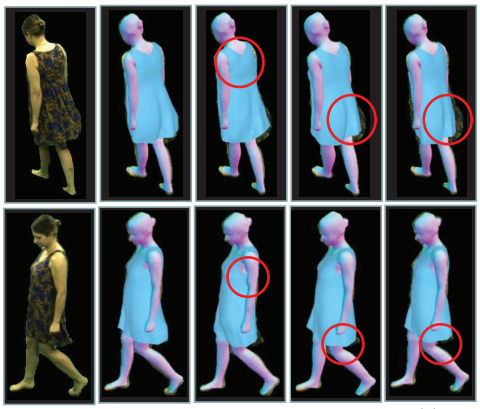
Figure 4. Comparison of novel-view synthesis with NerfCap [34]. (a,d) Ground truth. (b,e) NerfCap [34]. (c,f) Our method.

4.3. Ablation studies

The ablation experiments are performed to demonstrate the effectiveness of our framework on the S4 from DeepCap dataset [10]. The quantitative results are listed in Table 3.

Effectiveness of different losses. We first evaluate the effectiveness of the main losses in our optimization, i.e., the simulation loss, the mask loss, and the color loss. We report the accuracy by removing each of them, and show qualitative comparison examples in Figure 5. Without the simulation loss, the body-cloth collision becomes severe, showing that the simulation loss can preserve physically-plausible body-cloth interactions. Using the input human mask can

improve the alignment accuracy with the human silhouettes. The color loss of dynamic NeRF can further enhance the detailed deformation on the human surface by posing dense appearance constraints for the geometry tracking.



(a) (b) (c) (d) (e)

Figure 5. Ablation study for different losses. (a) Ground truth. (b) Using complete losses. (c) W/o simulation loss. (d) W/o mask loss. (e) W/o color loss. There are unnatural cloth deformations without the simulation loss. The rendered mask cannot align with the silhouette well without the mask loss. Compared to that without color loss, the surface detailed deformation can be further improved using complete losses (b).

(a) (b) (c)

Figure 6. Comparison of different rendering. (a) Ground truth. (b) Our composite rendering. (c) Blend rendering [6]. We show the synthesis of the clothing, body, and full human. There are much cloth appearance on the body caused by the blending [6], while our method separates the appearance of the clothing and body better.

Effectiveness of the composite rendering. We compare our composite rendering with the blending rendering in [6]. Through free composition of two NeRFs for each point, the method of [6] can obtain high fitting color. However, the blending may generate missing areas or artifacts in the novel-view synthesis. Moreover, it fails to recover the appearance of the body and clothing reliably as shown in Figure 6. Comparatively, we can recover the appearance of the body and clothing separately and compose them into a full human more accurately using the composite rendering.

The number of camera views. We investigate the influence of different number of input camera views on the performance of our approach. We compare the results of our method under $\{1, 2, 4, 6\}$ views, and show the comparison example of novel-view synthesis in Figure 7. The results become slightly better using more camera views because the moving humans are captured more completely.

Ground Truth 1 view 2 views 4 views 6 views

Figure 7. Ablation study for different numbers of camera views. With more cameras, the novel-view synthesis quality becomes better (marked in cycle).

4.4. Applications

Benefiting from separate tracking results of human body and clothing, we can enable interesting applications like cloth retargeting as shown in Figure 8. Since the body mesh and cloth mesh are captured, we can exchange the garments between two person by simulating the garments on the other body. Moreover, free-viewpoint motion can be created by simulating the clothing on human body sequences.

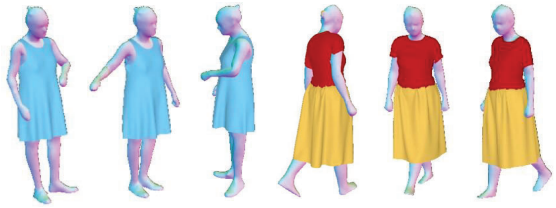


Figure 8. Application of cloth retargeting. From the captured results using our method, we can retarget the garments between the S4 of DeepCap and the 'FranziRed" of DynaCap.

5. Conclusion

In this paper, we proposed a novel performance capture method for clothed humans with a double-layer neural radiance fields (NeRF). Our double-layer NeRFs can effectively model moving humans in daily-life clothes. By matching the rendered color with the observed color, our method estimates the clothing and body motion and temporally-varying appearance robustly. In addition, we learn a differential physics-aware network for different garment types which supports physically-plausible body-cloth interactions and cloth dynamics. The experimental results on real datasets demonstrate the effectiveness of the proposed method.

Acknowledgments This work was supported by the Fundamental Research Funds for the Central Universities (NJ2022028), and the Natural Science Foundation of China (No.U1713208), and Program for Changjiang Scholars.

References

[1] https://www.marvelousdesigner.com/. 3, 5

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2Shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 2

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

[4] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. Fusion4D: Real-time performance capture of challenging scenes. *ACM TOG*, 35:1–13, July 2016. 1, 2

[5] Stuart Ganan and D McClure. Bayesian image analysis: An application to single photon emission tomography. In *Proceedings of the American Statistical Association*, pages 12–18, 1985. 5

[6] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, 2021. 4, 7, 8

[7] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. GarNet: A two-stream network for fast and accurate 3d cloth draping. In *ICCV*, 2019. 5

[8] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics*, 40(4), August 2021. 2, 5, 6, 7

[9] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. LiveCap: Real-time human performance capture from monocular video. *ACM TOG*, 38(2):14:1–14:17, March 2019. 1, 2

[10] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. DeepCap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 1, 2, 5, 6, 7

[11] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. SelfRecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. 2

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

[13] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 4

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015. 6

[15] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In *SIGGRAPH*, 2000. 3, 4

[16] Yue Li, Marc Habermann, Bernhard Thomaszewski, Stelian Coros, Thabo Beeler, and Christian Theobalt. Deep physics-aware inference of cloth deformation for monocular human performance capture. In *International Conference on 3D Vision (3DV)*, 2021. 1, 2, 5, 6

[17] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural Actor: Neural free-view synthesis of human actors with pose control. *ACM TOG*, 40(6):1–16, 2021. 3

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):248, 2015. 2, 3

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

[20] 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, 4

[21] Richard Newcombe, Dieter Fox, and Steve Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 1, 2

[22] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, 2021. 3, 4

[23] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In *CVPR*, 2020. 2, 5, 6, 7

[24] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 2, 3, 4, 5

[25] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2, 3, 4

[26] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM TOG*, 36(4), July 2017. 1, 2

[27] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 2

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

[29] Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. *Computer Graphics Forum (Proc. Eurographics)*, 2019. 2

[30] Vincent Sitzmann, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, 2020. 4, 6

[31] Zhaoqi Su, Weilin Wan, Tao Yu, Lingjie Liu, Lu Fang, Wenping Wang, and Yebin Liu. MulayCap: Multi-layer human performance capture using a monocular video camera. *IEEE Transactions on Visualization and Computer Graphics*, 28(4):1862–1879, 2022. 1, 2

[32] Robert W. Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM TOG*, 26(3), 2007. 3, 5

[33] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 2

[34] Kangkan Wang, Sida Peng, Xiaowei Zhou, Jian Yang, and Guofeng Zhang. NerfCap: Human performance capture with dynamic neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 1, 3, 4, 5, 6, 7

[35] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, pages 7467–7477, 2020. 5

[36] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit clothed humans obtained from normals. In *CVPR*, June 2022. 2, 6, 7

[37] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. MonoPerfCap: Human performance capture from monocular video. *ACM TOG*, 37(2):1–15, July 2018. 1, 2

[38] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera. In *ICCV*, October 2017. 1, 2

[39] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*, June 2018. 1, 2

[40] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. SimulCap: Single-view human performance capture with cloth simulation. In *CVPR*, 2019. 1, 2, 5

[41] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. HumanNeRF: Generalizable neural human radiance field from sparse inputs. In *CVPR*, 2022. 1, 3

[42] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D human reconstruction from a single image. In *ICCV*, 2019. 2

[43] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep Fashion3D: A dataset and benchmark for 3d garment reconstruction from single images. In *ECCV*, pages 512–530, 2020. 1, 2

[44] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering Explicit to Implicit: Towards high-fidelity garment mesh reconstruction from single images. In *CVPR*, pages 3845–3854, 2022. 1, 2

[45] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, 2019. 2