AniDress: Animatable Loose-Dressed Avatars from Sparse Views Using Garment Rigging Models

Beijia Chen*   Yuefan Shen*   Qing Shuai   Xiaowei Zhou   Kun Zhou   Youyi Zheng†

State Key Lab of CAD&CG, Zhejiang University

Abstract

Recent communities have seen significant progress in building photo-realistic animatable avatars from sparse multi-view videos. However, current workflows struggle to render realistic garment dynamics for loose-fitting characters as they predominantly rely on naked body models for human modeling while leaving the garment part unmodeled. This is mainly due to that the deformations yielded by loose garments are highly non-rigid, and capturing such deformations often requires dense views as supervision. In this paper, we introduce AniDress, a novel method for generating animatable human avatars in loose clothes using very sparse multi-view videos (4-8 in our setting). To allow the capturing and appearance learning of loose garments in such a situation, we employ a virtual bone-based garment rigging model obtained from physics-based simulation data. Such a model allows us to capture and render complex garment dynamics through a set of low-dimensional bone transformations. Technically, we develop a novel method for estimating temporal coherent garment dynamics from a sparse multi-view video. To build a realistic rendering for unseen garment status using coarse estimations, a pose-driven deformable neural radiance field conditioned on both body and garment motions is introduced, providing explicit control of both parts. At test time, the new garment poses can be captured from unseen situations, derived from a physics-based or neural network-based simulator to drive unseen garment dynamics. To evaluate our approach, we create a multi-view dataset that captures loose-dressed performers with diverse motions. Experiments show that our method is able to render natural garment dynamics that deviate highly from the body and generalize well to both unseen views and poses, surpassing the performance of existing methods. The code and data will be publicly available.

1. Introduction

Building animatable clothed human avatars has long been a challenging task in VR and AR. Recent neural rendering-

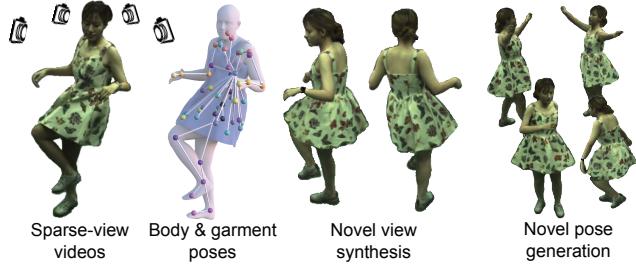*Equal Contributions, †Corresponding Author.

Figure 1. Given a sparse multi-view video with a loose-dressed performer, we estimate both body and garment poses aided by a garment rigging model. Then, a pose-driven neural radiance field is optimized to fit the video. At test time, our method can synthesize plausible body and garment motions from novel views. In this case, we use body motions from AMASS [34] and garment poses from physics-based simulation for novel pose synthesis.

based works [24, 27, 37, 41, 42, 60, 75, 76] enable us to learn animatable characters directly from sparse multi-view videos (usually consists of 300-400 frames) and produce photo-realistic rendering. Generally, these methods first capture the body movements using template-based estimators [1] and then draft the estimated deformations using a NeRF-based rendering model for appearance learning. At test time, these methods generate novel animations driven by unseen body poses.

Though appealing results have been achieved, these methods fail when applied to loose-fitting subjects. This is due to the fact that loose garments such as dresses and skirts, exhibiting complex non-rigid deformations such as swing and sliding effects, can own different movement patterns from body [32, 33, 39, 64, 65]. Using 3D body poses alone for capturing cross-frame correspondences is inaccurate, producing blurry renderings at training time. Moreover, these approaches generate novel animations driven by body pose only, failing to model the complex interplay between body and cloth dynamics at test time.

To address the above challenges, several methods have been introduced. HDHumans [14] exploits deformation graphs [53] for modeling garment movements. However,

their method requires additional annotations of ground-truth texture for each frame during training. Moreover, they learn the body-to-garment mapping directly from multi-view videos of limited length, resulting in stiff animations at test time. Other works [64, 65] track and reconstruct accurate garment meshes at the texture-aligned level for learning realistic rendering models. Novel garment statuses upon unseen body poses are obtained from physics-based simulation to render garment animations at test time. However, they rely on expensive dense views of cameras and cannot be applied to sparse views.

In this paper, we present *AniDress*, a novel solution for building generalizable loose-fitting avatars from a sparse multi-view video. Instead of directly learning geometry animation model and appearance model from limited video data, we follow the previous work [65] that only learns the rendering model from multi-view videos. To enable garment dynamics capture and rendering in very sparse views, we employ a garment rigging model built from physics-based simulation (PBS) data. Specifically, given a garment template (obtained from reconstruction), we first compute diverse garment dynamics for different body animations using PBS. Based on these mesh examples, a virtual bone-based garment rigging model is extracted via skinning decomposition [21]. This model encodes high-dimensional vertex deformations into a low-dimensional intrinsic garment space, empowering us with several advantages. First, it allows us to capture garment dynamics effectively using a set of bone transformations (denoted as garment poses for brevity). Technically, we introduce a novel method based on differentiable rendering to estimate temporally coherent garment poses from a multi-view video. Second, the estimated garment poses, in conjunction with body poses, constitute a unified signal to drive successive rendering. Specifically, we develop a pose-driven NeRF-based rendering method where both body and garment poses are used to deform a neural radiance field. Our method can generate photo-realistic avatars with controllability over the body and garment poses. At test time, garment poses can be derived from PBS, predicted from learning-based methods, or estimated from videos, to control the rendering of garment dynamics (see Fig. 1).

To evaluate our approach, we create a multi-view dataset capturing performers in diverse loose garments under various motions. Experiments show that our approach outperforms prior works in terms of both novel view and novel pose synthesis.

In summary, our contributions are as follows:
• We adopt a garment rigging model built from simulation data for capturing, animating, and rendering garment dynamics.
• A novel method to estimate garment poses from sparse multi-view RGB videos.
• A deformable neural radiance field conditioned on both garment and body poses, allowing for rendering high-quality body movements with natural garment dynamics.
• A multi-view dataset that captures five loose dresses in diverse motions for evaluation and further research.

2. Related Works

2.1. Clothed Body Avatars

Pioneer works learn animatable avatars from real scans [4, 26, 31–33, 47, 56, 71], exploring various geometry representations ranging from topology-fixed meshes [31] to topology-flexible forms such as point clouds [32, 33] and implicit functions [4, 35, 38, 47, 56]. However, these methods necessitate 3D body scans for training and additional rendering models for generating high-quality images.

Recent advances in neural rendering have facilitated learning animatable avatars directly from multi-view videos [5, 24, 41, 42, 44, 60]. Pioneer work [44] learns a NeRF-based rendering model in canonical space and drafts a deformation network for dynamic modeling, requiring dense view cameras for supervision. Subsequent research reduces the capture requirements from dense to sparse views [41, 42] or even to monocular setups [62, 69] and boosts its rendering speed from offline to real-time [5, 11, 74]. Additionally, geometry constraints [24, 60] and advanced dynamics learning schemes [62, 63, 69] have been investigated to develop more generalizable models for unseen poses.

However, existing methods primarily cater to avatars with tight clothing and face challenges when extending to loose garments, as they only model body dynamics. Xiang *et al.* [64–66] employ a dual-layer mesh representation to explicitly model garments. While capable of rendering high-quality images, these approaches demand accurate geometry reconstructions, requiring dense cameras or RGB-D inputs. Another family of research, e.g., [16, 18, 74],

Table 1. We compare the characteristics of our method with representative existing methods. Our method is the only one that can create loose-dressed avatars from sparse RGB views.

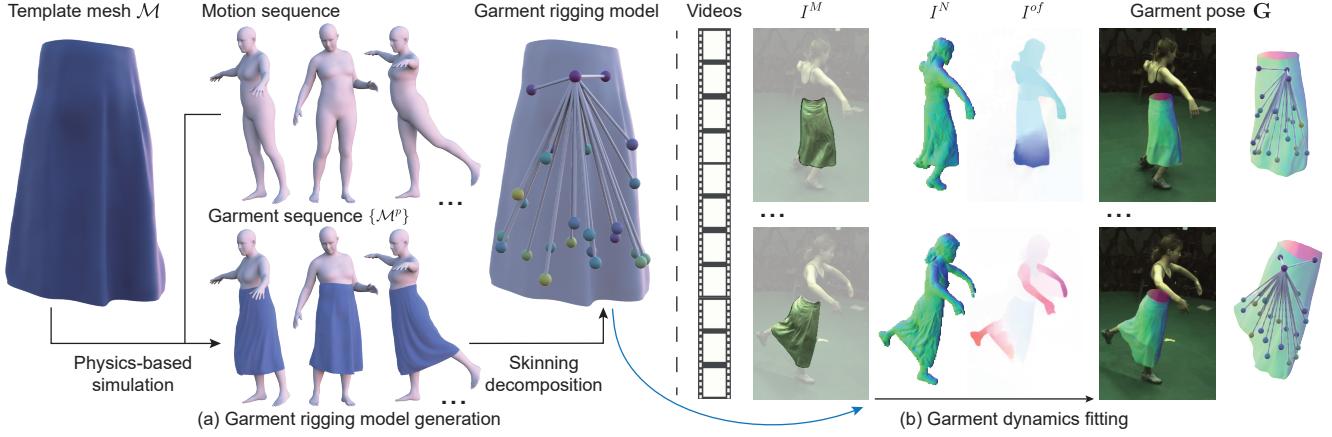| Methods	Sparse views	RGB only	Novel pose	Loose-dressed
NeuralBody [42]	✓	✓	✓	✗
Anim-Nerf [43]	✓	✓	✓	✗
Uv-Volumes [5]	✓	✓	✓	✗
Xiang *et al.* [65]	✗	✓	✓	✓
Xiang *et al.* [66]	✗	✗	✓	✓
Zhao *et al.* [74]	✗	✓	✗	✓
HumanRF [16]	✗	✓	✗	✓
FlexNeRF [18]	✓	✓	✗	✓
Ours	✓	✓	✓	✓

Figure 2. Overview of the procedures for building garment rigging model and capturing garment poses from a multi-view video. Starting from a template mesh $\mathcal{M}$, we run the physics-based simulation to generate diverse garment shapes $\{\mathcal{M}^p\}$, from which a garment LBS modeling is extracted via skinning decomposition. In the fitting step, we use the garment masks I^M, image normals I^N, and optical flows I^{of} to estimate the garment poses at each frame.

focuses on capturing human performance with diverse garments using multi-view cameras but lacks the ability to generalize to unseen poses.

In contrast, our method relies solely on sparse multi-view videos as input and demonstrates the capacity to generalize garment dynamics under novel poses. The characteristics of our method and several representative existing methods are summarized in Tab. 1.

2.2. Garment Capture

Capturing 3D garments is essential for various applications, including virtual try-ons and generative modeling. Several methods have been developed to reconstruct static garments from a single image [19, 52, 77]. Recent works [10, 45] further enable us to capture dynamic garments from a monocular video. However, monocular videos inevitably introduce shape ambiguity for garment reconstruction and the temporal coherency of these methods can not be ensured. Our research, however, focuses on recovering temporally coherent 3D garments for appearance learning from a sparse multi-view video, differing from these approaches.

Recent works from Xiang *et al.* [64–66] rely on well-reconstructed garment geometries under dense views and register template garments to reconstructions using non-rigid ICP (Iterative Closest Point) [22]. [12, 14, 25] present approaches for predicting garment deformation from a single image, achieved by training a manually defined deformation graph to align with a pre-captured multi-view dataset. In contrast, we only have sparse views of short videos as input, and it is difficult to reconstruct the garment geometry accurately or pre-train a deformation-predicting network.

2.3. Garment Animation

Modeling realistic garment dynamics driven by non-rigid body motions is crucial for creating high-fidelity human avatars. Research in this domain generally falls into two main categories: physics-based simulation [2, 29, 54, 57, 70] and data-driven methods [3, 6–8, 15, 20, 39, 40, 48, 58, 61, 72]. Here, we focus on a recent study closely related to our work. Instead of directly inferring animations at the mesh level, [39] propose to transfer body motions to a set of bone transformations of a pre-computed garment skinning model. However, their research primarily investigates the use of this skinning model for animating garment geometry for faster simulation. In our paper, we delve deeper into the challenges of capturing and rendering garment dynamics in real-captured videos utilizing a similar model. It is important to note that their approach is complementary to ours, as it provides a foundation for inferring garment bone motions for unseen body movements at test time.

3. Method

Our method takes as input a sparse multi-view video of a performer wearing a loose garment along with the corresponding garment template mesh. Our goal is to construct an animatable loose-fitting avatar that supports photo-realistic rendering of not only body poses but also garment dynamics.

To achieve so, our method consists of three major steps. First, we build a garment rigging model from offline physics-based simulation data, encoding possible garment shapes into low-dimensional garment poses (Sec. 3.1). Then, a novel garment fitting method is introduced for estimating the coarse garment dynamics from sparse multi-view videos $\mathcal{I}$ aided by such a rigging model. Please see

Fig. 2 for an overview of the above two steps. Being equipped with both body and garment poses, we learn a generalizable pose-driven rendering network from $\mathcal{I}$, where both garment and body poses are exploited to drive a deformable neural radiance field (Sec. 3.3 and Fig. 3). At test time, our method requires both body and garment poses to produce realistic rendering. We show how garment poses can be obtained for unseen body poses in Sec. 3.4.

3.1. Garment Rigging Model

In this subsection, we first introduce the formulation of our garment rigging model and then show how to build it from simulation data. We adopt Linear Blend Skinning (LBS) for our garment rigging model. In the LBS model, mesh deformations are driven by a set of bones. Given a template mesh $\mathcal{M}$ in rest pose with $\mathbf{v}_i$ denotes the position of its i-th vertex, the deformed i-th vertex $\mathbf{v}_i^p$ at configure p can be represented as:

$$\mathbf{v}_i^p = \sum_{j=1}^{B} w_{ij} \mathbf{T}_j^p \mathbf{v}_i \qquad (1)$$

where $\mathbf{T}_j^p$ represents j-th bone's transformation matrix at configure p, and B is the number of bones. $\mathbf{W} = \{w_{ij}\}$ is the mesh skinning weights and each scalar w_{ij} reflects the influence of j-th bone to the i-th vertex and satisfies $w_{ij} \geq 0$ and $\sum_{j=1}^{B} w_{ij} = 1$. At configure p, we refer $\mathbf{G}^p = \{\mathbf{T}_j^p\}_{j=1}^{B}$ as garment poses and denote the deformed mesh $\mathcal{M}^p = \text{LBS}(\mathcal{M}, \mathbf{W}, \mathbf{G}^p)$ for compactness.

Given a set of deformed meshes $\{\mathcal{M}^p\}$, learning such an LBS model equals localizing B bones in canonical space and solving the joint transformations $\mathbf{G}^p$, together with skinning weight matrix $\mathbf{W}$, that best explain the observations $\{\mathcal{M}^p\}$. To ensure the capacity of such LBS representation, we propose to learn it from diverse garment examples obtained from simulation.

Physics-based Garment Simulation. Specifically, we select several body motions with high dynamics covering walking, running, and dancing from the AMASS dataset [34]. Together with the captured body motions from our multi-view video dataset, we then run the physics-based simulation to obtain diverse garment shapes $\{\mathcal{M}^p\}$, as shown in Fig. 2.

Skinning Decomposition. Given the simulated mesh sequences, we further encode the garment dynamics from high-dimensional vertex space into an LBS model using the state-of-the-art skinning decomposition method Smooth Skinning Decomposition with Rigid Bones (SSDR) [21]. Specifically, SSDR solves bone transformations $\mathbf{G}^p$ and skinning weights $\mathbf{W}$ by minimizing the mesh reconstruction loss. We refer the readers to [21] for more details about decomposition. Following [39], we refer to the extracted skeleton as virtual bones since these joints may not actually located at mesh surfaces. Such an LBS model forms a com-

pact representation of the underlying geometry, enabling us to capture and animate garment dynamics via garment poses rather than high-dimensional vertices.

3.2. Garment Dynamics Fitting

Being equipped with the rigging model, we are now able to estimate garment dynamics from sparse multi-view videos using bone transformations. Formally, we estimate garment poses $\mathbf{G}^t$ for each timestep t, while ensuring the alignments between garment meshes and sparse view observations $\mathcal{I}$.

However, such a task still poses significant challenges due to the inherent complexities of fabric behavior. Garment deformations, which are complex and non-linear, construct a large space of valid shapes. Capturing folds and drapes of garments, which constantly change with body movements, is ill-posed. Unlike body mesh fitting where explicit keypoints detection and kinematic constraints are often used, limited regularizations can be used for estimating garment dynamics. Moreover, ensuring temporal consistency across frames during estimation is also challenging [45].

To address the above challenges, we develop a novel optimization method based on differentiable rendering [28, 46] where we align 3D garment shapes to a set of robust 2D cues, i.e., garment silhouettes, image normals, and optical flows. We show each of these cues boosts fitting performance at one specific aspect and is essential for obtaining valid garment dynamics (see Sec. 5.3).

To begin with, we optimize the difference between rendered garment masks and detected 2D silhouettes. However, silhouette loss can only offer a matched 3D outer shape, leading to invalid folds and drapes. To predict more accurate folds, we further incorporate normal cues predicted from 2D images. To enhance the temporal consistency across frames, mesh deformations are projected in 2D space to match pixel movements of optical flow. The right part of Fig. 2 illustrates our fitting step.

Overall, the optimization problem can be formulated as:

$$\begin{aligned} \underset{\mathbf{G}^t}{\text{minimize}} & \, \mathcal{L}_{fit}(\text{LBS}(\mathcal{M}, \mathbf{W}, \mathbf{G}^t)), \\ \mathcal{L}_{fit} & = \lambda_m \mathcal{L}_M + \lambda_n \mathcal{L}_N + \lambda_{of} \mathcal{L}_{of}. \end{aligned} \qquad (2)$$

where $\mathcal{L}$ are different terms of loss terms and λ represent their weights. We illustrate each loss term individually in the following.

Garment Silhouette Loss. We use [23] to get garment silhouettes I^M and silhouette loss in multi-views is defined as:

$$\mathcal{L}_M = \sum_{v}^{N^v} ||\mathcal{DR}(\text{LBS}(\mathcal{M}, \mathbf{W}, \mathbf{G}^t)), c_v) - I_v^M||_2^2, \qquad (3)$$

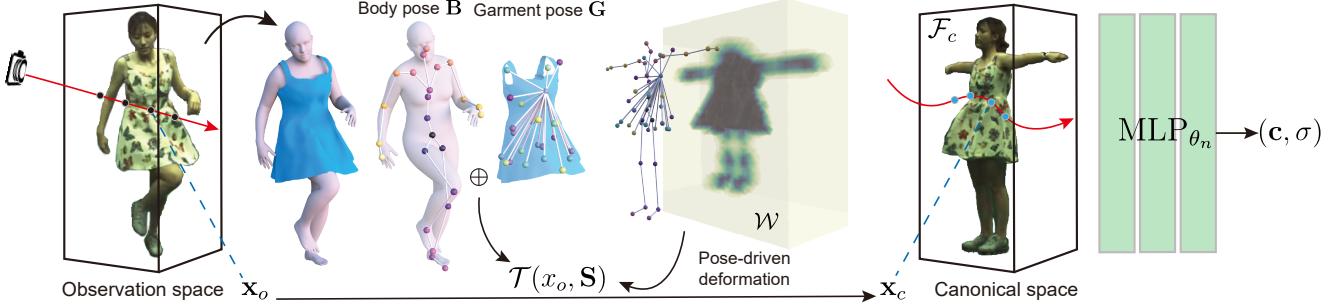where $\mathcal{DR}$ is the differentiable renderer [28], N^v is the

Figure 3. Overview of our rendering pipeline. For each sampled point in observation space, we first transform it back to the canonical space using a pose-driven deformation module conditioned on both body $\mathbf{B}$ and garment poses $\mathbf{G}$ and then query its color and density $(\mathbf{c}, \sigma)$ through a radiance field $\mathcal{F}_c$ defined in the canonical space.

number of views, and c_v is the camera parameters of the v-th view.

Image Normal Loss. We use normal predictor proposed by [67] to estimate the normal map I^N from 2D images and formulate the normal projection loss as:

$$\mathcal{L}_N = \sum_v^{N^v} ||\mathcal{N}^p(\text{LBS}(\mathcal{M}, \mathbf{W}, \mathbf{G}^t)), c_v) - I_v^N * I_v^M||_2^2, \quad (4)$$

where $\mathcal{N}^p(\cdot, c)$ is the rendered normals of the visible mesh faces in view c. We also multiple the garment mask I^M on the predicted normal map.

Optical Flow loss. Once we get the fitted geometry $\mathcal{M}^{t-1}$ of $(t\text{-}1)$-th frame, we can use it as the initialization for the next timestep and leverage the predicted 2D optical flow I^{of} as temporal constraint between frames. We employ the commonly-used RAFT [55] to estimate the optical flow between two adjacent frames, and the optical flow loss is defined as:

$$\begin{aligned} \mathcal{L}_{of} = \sum_v^{N^v} ||\mathcal{C}^p(\text{LBS}(\mathcal{M}, \mathbf{W}, \mathbf{G}^t), c_v) \\ - (\mathcal{C}^p(\mathcal{M}^{t-1}, c_v) + \delta_v)||_2^2, \\ \delta_v = I_v^{of}(\mathcal{C}^p(\mathcal{M}^{t-1}, c_v)), \end{aligned} \quad (5)$$

where $\mathcal{C}^p(\cdot, c)$ means visible projected coordinates in the screen space in the camera view c. We optimize the projected mesh movement between $\mathcal{M}^{t-1}$ and the newly predicted shape to match the optical flow value δ in I^{of}. Notably, for the first frame, we only use $\mathcal{L}_M$ and $\mathcal{L}_N$ because there is no previous frame for estimating optical flow.

3.3. Pose-driven Rendering Network

Given a sparse multi-view video $\mathcal{I}$, along with the estimated body and garment poses, we aim to build a generalizable NeRF-based rendering module with controllability over both body and garment dynamics.

Following previous works [41], we represent a moving person with a neural radiance field $\mathcal{F}_c$ in the canonical space. A deformation module $\mathcal{T}$ is used to map points from observation space back to the canonical space for efficient rendering. The challenge here is how one learn such deformation accurately. Previous template-based methods [41, 42] drives near-surface points using mesh deformations. However, simply applying such strategy for modeling both body and garment movements leads to sub-optimal performance as it requires us to not only estimate accurate geometries for both parts but also resolve the body-garment mesh penetrations.

To address the above challenges, we propose to learn a pose-driven deformation field conditioned on both 3D body and garment poses. Specifically, the color and density of point $\mathbf{x}_o$ in the observation space is defined as:

$$(\mathbf{c}(\mathbf{x}_o), \sigma(\mathbf{x}_o)) = \mathcal{F}_c(\mathcal{T}(\mathbf{x}_o, \mathbf{G}, \mathbf{B})), \quad (6)$$

where $\mathcal{T}$ is a backward deformation module that takes both body pose $\mathbf{B}$ and garment pose $\mathbf{G}$ as input, and outputs the point $\mathbf{x}_c$ in canonical space. $\mathcal{F}_c$ is a multi-layer perceptrons MLP_{θ_n} that takes point in the canonical space $\mathbf{x}_c$ as input and outputs its color value $\mathbf{c}$ and density σ.

Pose-driven Deformation Module. Instead of modeling body and garment motions independently, which requires us to further infer their interactions, we concatenate them together to form a unified pose $\mathbf{S}$, as shown in Fig. 3. Then, the transformation $\mathcal{T}$ is defined as an inverse LBS function that deforms points in the observation space to canonical space:

$$\mathcal{T}(\mathbf{x}_o, \mathbf{S}) = \sum_k^J w_o^k \mathbf{T}_k^{-1} \mathbf{x}_o, \quad (7)$$

where $\mathbf{T}_k^{-1}$ is the inverse transformation matrix of k-th joint of $\mathbf{S}$, J is the number of joints and w_o^k is inverse blend weights of $\mathbf{x}_o$ respects to k-th joint. However, this formulation requires us to solve each observation space an inverse blend weight field, which is computation intensive.

5

Moreover, the learned model may be over-fitted to training frames and generalizes poorly to unseen poses.

To address the above challenge, we follow [62] to define the inverse blend weight field in one shared canonical space. Specifically, we use a CNN_{θ_t} parameterized by θ_t to generate the inverse blend weight field $\mathcal{W}$ from a fixed random latent code. Then w_o^k is computed as:

$$w_o^k = \frac{w_c^k \mathbf{T}_i^{-1} \mathbf{x}_o}{\sum_{i=1}^{J} w_c^i \mathbf{T}_i^{-1} \mathbf{x}_o}, \quad (8)$$

where $w_c^k = \mathcal{W}(\mathbf{T}_i^{-1} \mathbf{x}_o)$ denotes the queried value in the blend weight field $\mathcal{W}$ at the position $\mathbf{T}_i^{-1} \mathbf{x}_o$ in the canonical space.

NeRF Optimization. Based on the above NeRF representation, we are able to optimize the canonical appearance module $\mathcal{F}_c$, together with the inverse blend weights volume $\mathcal{W}$ using volume rendering techniques [36]. Specifically, we cast a ray $\mathbf{r}$ at observation space and minimize the mean squared error (MSE) between the rendered RGB color $\tilde{\mathbf{C}}$ with the ground truth $\mathbf{C}$:

$$\mathcal{L}_{\text{MSE}} = \sum_{i \in N^f} \sum_{j \in N^v} \sum_{r \in \mathcal{R}} ||\tilde{\mathbf{C}}_{i,j}(\mathbf{r}) - \mathbf{C}_{i,j}(\mathbf{r})||_2^2, \quad (9)$$

where N^f, N^v are numbers of frames and views, and $\mathcal{R}$ is the set of rays passing through images pixels. We also adopt a perceptual loss LPIPS [73] and the final loss of our rendering network is $\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{LPIPS}}$.

3.4. Test Time Animation

At test time, our method requires both novel body and garment poses to produce photo-realistic rendering. Here, we show several ways to obtain valid garment poses. First, novel garment poses can be estimated from a newly-come multi-view video, as shown in Fig. 6. In such a situation, garment poses serve as control signals to reproduce the garment dynamics presented in the newly-come multi-video under arbitrary viewpoints without retraining the rendering model. Second, garment poses can be borrowed from our simulated and decomposed garment sequences to drive unseen dynamics, shown in Sec. 5.4. Thirdly, we can run the physics-based simulation to obtain new garment shapes, and the garment poses can be further obtained via mesh fitting. Moreover, as discussed in Sec. 2.3, existing work [39] has already explored the possibility of inferring garment poses for unseen body poses using neural networks and achieved impressive results. Their work motivates us and serves as a strong complementary as they mainly focus on animating garment poses for novel body motion and our method creates realistic garment rendering models from real data.
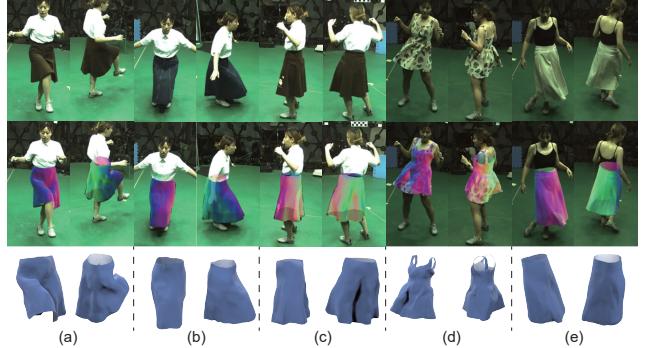

Figure 4. Visualizations of our garment fitting results, where the 1st and the 3rd rows show the inputs and rendered garment geometry, respectively. We also overlay the rendered mesh geometry onto input images to illustrate the alignments (the 2nd row).

4. Dataset

The commonly used multi-view video datasets for human modeling, such as H3.6M [17] and ZJU-MoCap [42], focus on subjects in tight clothing and are thus unsuitable for evaluating our method. We create a new dataset comprising 12 dynamic videos featuring a performer in 5 different loose clothing such as dresses and skirts. We captured at least 14 camera views for every motion, selecting eight uniformly distributed cameras for training and the rest for testing novel view performance. Each garment is represented in at least two sequences showcasing varied body movements to facilitate evaluation in novel poses. When we exploit one of the videos for training, the other video will be used as test data for novel pose evaluation. The length of each sequence ranges from 400 to 800 frames. For each sequence, we generate binary masks for the human body and garment using [23] and estimate 3D body poses with [1]. The mesh template for each garment is derived from static reconstruction, and more details can be found in our supplementary material.

5. Experiments

5.1. Results of Garment Fitting

In Fig. 4, a collection of garment fitting results is presented, demonstrating the efficacy of the novel fitting method proposed in Sec. 3.2. Though our method fails to capture details such as wrinkles, the fitted garment meshes align with the input images in silhouettes and exhibit coarse folds similar to observations (Fig. 4 (a) and (c)). Our method also successfully captures dynamics such as skirt swings (Fig. 4 (b) (d) and (e)). The recovered dynamics is temporally coherent and we recommend the readers to see our supplementary video for more visualization.

Figure 5. Visual comparisons on novel view synthesis. We also show an extra view of our results denoted as Ours′ and the GT′.

NeuralBody   AnimatableNerf   UV-Volumes   w/o garment pose   Ours   GT   Ours′   GT′
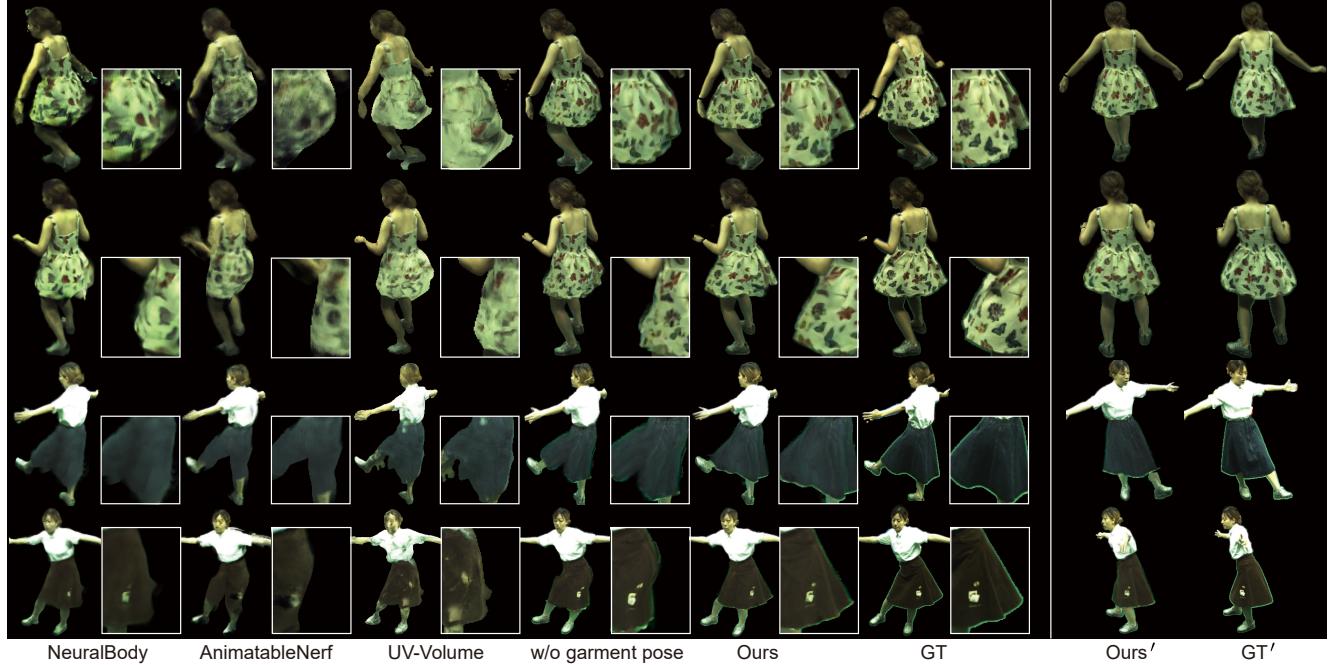

Figure 6. Visual comparisons on novel pose generation. We also show an extra view of our results denoted as Ours′ and the GT′.

NeuralBody   AnimatableNerf   UV-Volume   w/o garment pose   Ours   GT   Ours′   GT′

Table 2. Quantitative comparisons on novel view synthesis. We color cells that have the best and second best scores. LPIPS∗ = LPIPS × 10, DISTS∗ = DISTS × 10.

	PSNR ↑	IoU ↑	LPIPS∗ ↓	DISTS∗ ↓
NeuralBody [42]	28.45	78.95	0.4239	1.1132
Anim-Nerf [41]	26.67	72.13	0.5866	1.5989
UV-Volumes [5]	26.82	88.81	0.3231	1.0078
Ours w/o **G**	27.12	85.27	0.3026	0.9208
Ours	27.40	86.14	0.2846	0.8838

Table 3. Quantitative comparisons on novel pose synthesis. We color cells that have the best and second best scores. LPIPS∗ = LPIPS × 10, DISTS∗ = DISTS × 10.

	PSNR ↑	IoU ↑	LPIPS∗ ↓	DISTS∗ ↓
NeuralBody [42]	25.31	75.97	0.5486	1.4317
Anim-Nerf [41]	24.70	71.19	0.6458	1.7347
UV-Volumes [5]	24.39	82.53	0.4899	1.4309
Ours w/o **G**	24.97	80.95	0.4095	1.1275
Ours	25.45	82.23	0.3774	1.0744

5.2. Image Quality Evaluation

Baselines: To validate our method, we compare it against several state-of-the-art human modeling methods: Neural-Body [42], Animatable-NeRF [41] and UV-Volumes [5]. NeuralBody [42] anchors feature vectors on a deformable mesh template and generates the radiance field using forward skinning. Animatable-NeRF [41] learn a backward neural blend weights field conditioned on 3D body poses and points are transformed back to canonical space for rendering. Different from previous methods [41, 42] that predict pixel colors directly, UV-Volumes [5] render UV coordinates of each pixel and the RGB values are then queried from a neural texture stack. We also compare our method with a variant that does not use our proposed garment rigging model, i.e., using body poses **B** only in the pose-driven rendering network, which is similar to a multi-view variant of [62].

Metrics: In line with [5], our evaluation employs the Peak Signal-to-Noise Ratio (PSNR) for pixel-level measurement and Learned Perceptual Image Patch Similarity (LPIPS) for patch feature-level assessment. Additionally, we utilize Intersection over Union (IoU) to assess the quality of the generated avatar's outer shape. Furthermore, we utilize Deep Image Structure and Texture Similarity (DISTS) [9] as a metric for evaluating texture similarity.

Novel View Synthesis: Fig. 5 shows the visual comparisons with baselines [5, 41, 42] and our variant without garment pose **G**. Baselines struggle to produce clear textures and accurate shapes for dynamic dresses. Benefiting from the estimated garment pose, our method successfully maintains consistency in both garment appearances and geometries across views. Quantitative comparisons are de-
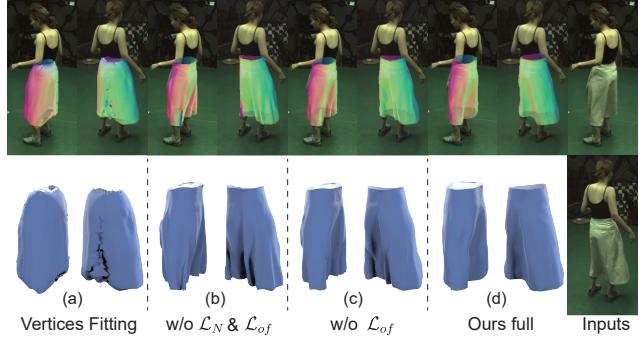

Figure 7. Visual comparisons of fitted mesh geometries. We compare our full method with (a) directly fitting mesh vertices, (b) without using normal $\mathcal{L}_N$ and optical flow $\mathcal{L}_{of}$ losses, and (c) without using the optical flow loss.

tailed in Tab. 2. LPIPS scores indicate the superiority of our method in generating textures that agree with human perceptions and DISTS scores show our ability to generate similar textures with the the ground truth. Furthermore, IoU scores demonstrate that our method achieves significantly more accurate outer shapes compared to [42]. Note that, UV-Volume [5] additionally uses a silhouette loss for training, thus generating accurate outer shapes. However, their textures do not match the quality of ours.

Novel Pose Generation: We estimate body and garment poses from the test motion sequences and drive trained avatars to generate novel poses for our method. Visual comparisons are shown in Fig. 6 and quantitative results are presented in Tab. 3. Existing methods like NeuralBody and UV-Volumes show limited generalizability with loose garments, while our method exhibits superior visual quality, featuring clear textures and correct shapes. It is noteworthy that, without the proposed garment rigging model, the generated dresses will have pants-like shapes, as shown in 3rd row Fig. 6. The superiority shown quantitatively further demonstrates the generalizability of our method.

5.3. Ablation Study

We provide ablation studies for the proposed garment fitting step to validate the robustness of our method. As shown in Fig. 7, comparisons between our full model and its ablated versions (a)-(c) highlight the efficacy of the introduced components. Ablation (a) abandons our garment rigging model and directly uses garment mesh vertices for fitting, resulting in invalid mesh geometries with serve self-penetrations. In ablation (b), both normal loss and optical flow loss are disabled. While the outer shape of the garment mesh aligns with the 2D silhouettes, the dress folds do not match image observations. In ablation (c), we remove optical flow loss. Using normal loss only, the optimization for the geometry shape simply deforms the vertices to meet the
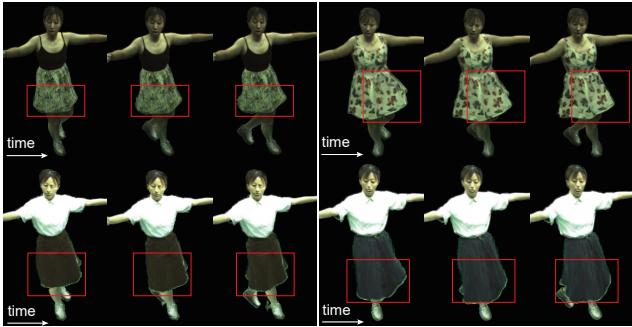
Figure 8. The generated renderings under novel pose sequences with garment poses driven by physics-based simulations. Please see the highlights for the dynamic differences.

normal prediction, resulting in garment self-penetrations. In contrast, our full model recovers the desired garment dynamics under the constraints of predicted optical flow. The incorporation of optical flow loss $\mathcal{L}_{of}$ enhances the temporal consistency of the fitted geometry sequence and we show this improvement in the supplementary video.

5.4. Novel Pose Generation Using Simulation Data

To further show the generalization ability of our rendering method, we generate images driven by garment poses produced by the physics-based simulation. Specifically, we select one certain dance body motion as well as four dress geometries for the simulation and compute the garment pose for each dress. Using the garment poses together with the body pose, we drive four pre-trained avatars to generate novel pose sequences. As shown in Fig. 8, our method can generate plausible dynamics for loose garments for unseen motions.

6. Discussion

The influence of physics-based simulation parameters. One of the most important steps in the proposed method is to run a stable and valid physics-based simulation to constitute a garment dataset. However, in-valid simulation parameters may lead to simulation failures. To address this problem, we conduct the simulation on each garment using Houdini Vellum [51] with default settings except for the fabric's bending stiffness and the simulator's time scale since these two parameters influence the simulated results the most. The time scale is adjusted based on their respective motion capture FPS. While the bending stiffness controls how soft the material is. We choose the bending stiffness to be 1e-4 for all garments since it corresponds to a soft material. This is based on the observations that using soft materials in simulation can make sure that the garment rigging model is able to express stiff status in capturing. However, if we start with a stiff material in simulation (i.e., larger bend stiffness), the garment rigging model may not

be able to capture soft garment dynamics.

Conclusion. We introduced *AniDress*, an innovative approach for building expressive loose-dressed animatable avatars, capable of synthesizing images in novel views and novel poses. By leveraging a garment rigging model, our method captures and renders garment dynamics using a set of bone transformations. Technically, we introduce a novel method for estimating garment poses to sparse multi-view videos, aided by 2D cues. To provide controllability over both body and garment dynamics, a pose-driven neural radiance field conditioned on both body and garment parts is introduced to render high-quality images. We build a new dataset consisting of loose-dressed performers in diverse body motions and demonstrate the superior performance of our method over baselines via extensive experiments.

Limitations and Future Works. Our method requires a template mesh for building the rigging model. So far, obtaining such template mesh requires manual correction. In the future, we will explore the possibility of automatic template mesh reconstruction. Moreover, our method relies on the garment rigging model built from simulation data. So far, the rigging model needs to be rebuilt for different garment types. In the future, we will explore the possibility of building a generalizable rigging model that can be applied to different garment types. Besides, we build our rendering module upon volume rendering where points far away from the garment surface may also contribute to the rendering. This may cause ghosting effects in challenging novel pose scenarios. Geometry constraints can be used to further regularize the NeRF optimization process.

References

[1] Easymocap - make human motion capture easier. Github, 2021. 1, 6

[2] David Baraff and Andrew Witkin. Large steps in cloth simulation. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 43–54, 1998. 3

[3] Hugo Bertiche, Meysam Madadi, Emilio Tylson, and Sergio Escalera. Deepsd: Automatic deep skinning and pose space deformation for 3d garment animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5471–5480, 2021. 3

[4] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 2

[5] Yue Chen, Xuan Wang, Xingyu Chen, Qi Zhang, Xiaoyu Li, Yu Guo, Jue Wang, and Fei Wang. Uv volumes for real-time rendering of editable free-view human performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16621–16631, 2023. 2, 8, 3, 4

[6] Enric Corona, Albrt Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11875–11885, 2021. 3

[7] Y D. Li, Min Tang, Yun Yang, Zi Huang, R F. Tong, SC Yang, Yao Li, and Dinesh Manocha. N-cloth: Predicting 3d cloth deformation with mesh-based networks. In *Computer Graphics Forum*, pages 547–558. Wiley Online Library, 2022.

[8] Edilson De Aguiar, Leonid Sigal, Adrien Treuille, and Jessica K Hodgins. Stable spaces for real-time clothing. *ACM Transactions on Graphics (TOG)*, 29(4):1–9, 2010. 3

[9] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 8, 3, 4

[10] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[11] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8770, 2023. 2

[12] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 3

[13] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)*, 40(4):1–16, 2021. 1

[14] Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–23, 2023. 1, 3

[15] Daniel Holden, Bang Chi Duong, Sayantan Datta, and Derek Nowrouzezahrai. Subspace neural physics: Fast data-driven interactive simulation. In *Proceedings of the 18th annual ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 1–12, 2019. 3

[16] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Trans. Graph.*, 42(4), 2023. 2

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 6

[18] Vinoj Jayasundara, Amit Agrawal, Nicolas Heron, Abhinav Shrivastava, and Larry S Davis. Flexnerf: Photorealistic free-viewpoint rendering of moving humans from sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21118–21127, 2023. 2

[19] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, pages 18–35, 2020. 3

[20] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European conference on computer vision (ECCV)*, pages 667–684, 2018. 3

[21] Binh Huy Le and Zhigang Deng. Smooth skinning decomposition with rigid bones. *ACM Transactions on Graphics (TOG)*, 31(6):1–10, 2012. 2, 4, 1

[22] Hao Li, Robert W Sumner, and Mark Pauly. Global correspondence optimization for non-rigid registration of depth scans. In *Computer graphics forum*, pages 1421–1430. Wiley Online Library, 2008. 3

[23] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 4, 6, 1

[24] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, pages 419–436. Springer, 2022. 1, 2

[25] 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 *2021 International Conference on 3D Vision (3DV)*, pages 373–384. IEEE, 2021. 3

[26] Zhe Li, Zerong Zheng, Hongwen Zhang, Chaonan Ji, and Yebin Liu. Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In *European Conference on Computer Vision*, pages 322–341. Springer, 2022. 2

[27] 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 transactions on graphics (TOG)*, 40(6):1–16, 2021. 1

[28] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[29] Tiantian Liu, Sofien Bouaziz, and Ladislav Kavan. Quasi-newton methods for real-time simulation of hyperelastic materials. *Acm Transactions on Graphics (TOG)*, 36(3):1–16, 2017. 3

[30] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 1

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 2

[32] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10974–10984, 2021. 1, 2

[33] Qianli Ma, Jinlong Yang, Michael J Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *2022 International Conference on 3D Vision (3DV)*, pages 679–689. IEEE, 2022. 1, 2

[34] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 4

[35] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10461–10471, 2021. 2

[36] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 2020. 6

[37] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5762–5772, 2021. 1

[38] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12695–12705, 2021. 2

[39] Xiaoyu Pan, Jiaming Mai, Xinwei Jiang, Dongxue Tang, Jingxiang Li, Tianjia Shao, Kun Zhou, Xiaogang Jin, and Dinesh Manocha. Predicting loose-fitting garment deformations using bone-driven motion networks. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 1, 3, 4, 6

[40] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7365–7375, 2020. 3

[41] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 1, 2, 5, 8, 3, 4

[42] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 1, 2, 5, 6, 8, 3, 4

[43] Sida Peng, Zhen Xu, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable implicit neural representations for creating realistic avatars from videos. *arXiv preprint arXiv:2203.08133*, 2022. 2

[44] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[45] Lingteng Qiu, Guanying Chen, Jiapeng Zhou, Mutian Xu, Junle Wang, and Xiaoguang Han. Rec-mv: Reconstructing 3d dynamic cloth from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4637–4646, 2023. 3, 4

[46] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 4

[47] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 2

[48] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, pages 355–366. Wiley Online Library, 2019. 3

[49] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[50] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[51] SideFX. Houdini vellum, 2021. 9, 1

[52] Astitva Srivastava, Chandradeep Pokhariya, Sai Sagar Jinka, and Avinash Sharma. xcloth: Extracting template-free textured 3d clothes from a monocular image. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2504–2512, 2022. 3

[53] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pages 80–es. 2007. 1

[54] Min Tang, Ruofeng Tong, Rahul Narain, Chang Meng, and Dinesh Manocha. A gpu-based streaming algorithm for high-resolution cloth simulation. In *Computer Graphics Forum*, pages 21–30. Wiley Online Library, 2013. 3

[55] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 5, 2

[56] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[57] Tzvetomir Vassilev, Bernhard Spanlang, and Yiorgos Chrysanthou. Fast cloth animation on walking avatars. In *Computer Graphics Forum*, pages 260–267. Wiley Online Library, 2001. 3

11

[58] Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. Fully convolutional graph neural networks for parametric virtual try-on. In *Computer Graphics Forum*, pages 145–156. Wiley Online Library, 2020. 3

[59] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1

[60] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European conference on computer vision*, pages 1–19. Springer, 2022. 1, 2

[61] Tuanfeng Y Wang, Tianjia Shao, Kai Fu, and Niloy J Mitra. Learning an intrinsic garment space for interactive authoring of garment animation. *ACM Transactions on Graphics (TOG)*, 38(6):1–12, 2019. 3

[62] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 2, 6, 8, 1

[63] Chung-Yi Weng, Pratul P Srinivasan, Brian Curless, and Ira Kemelmacher-Shlizerman. Personnerf: Personalized reconstruction from photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 524–533, 2023. 2

[64] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)*, 40(6):1–15, 2021. 1, 2, 3

[65] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022. 1, 2

[66] Donglai Xiang, Fabian Prada, Zhe Cao, Kaiwen Guo, Chenglei Wu, Jessica Hodgins, and Timur Bagautdinov. Drivable avatar clothing: Faithful full-body telepresence with dynamic clothing driven by sparse rgb-d input. In *ACM SIGGRAPH Asia 2023 Conference Proceedings*, pages 1–9, 2023. 2, 3

[67] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 5, 2

[68] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[69] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 2

[70] Cyril Zeller. Cloth simulation on the gpu. In *ACM SIGGRAPH 2005 Sketches*, pages 39–es. ACM New York, NY, USA, 2005. 3

[71] Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. Closet: Modeling clothed humans on continuous surface with explicit template decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 501–511, 2023. 2

[72] Meng Zhang, Tuanfeng Y Wang, Duygu Ceylan, and Niloy J Mitra. Dynamic neural garments. *ACM Transactions on Graphics (TOG)*, 40(6):1–15, 2021. 3

[73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6, 3, 4

[74] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, et al. Human performance modeling and rendering via neural animated mesh. *ACM Transactions on Graphics (TOG)*, 41(6):1–17, 2022. 2

[75] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022. 1

[76] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive fullbody avatars. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 1

[77] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 512–530. Springer, 2020. 3

AniDress: Animatable Loose-Dressed Avatars from Sparse Views Using Garment Rigging Models
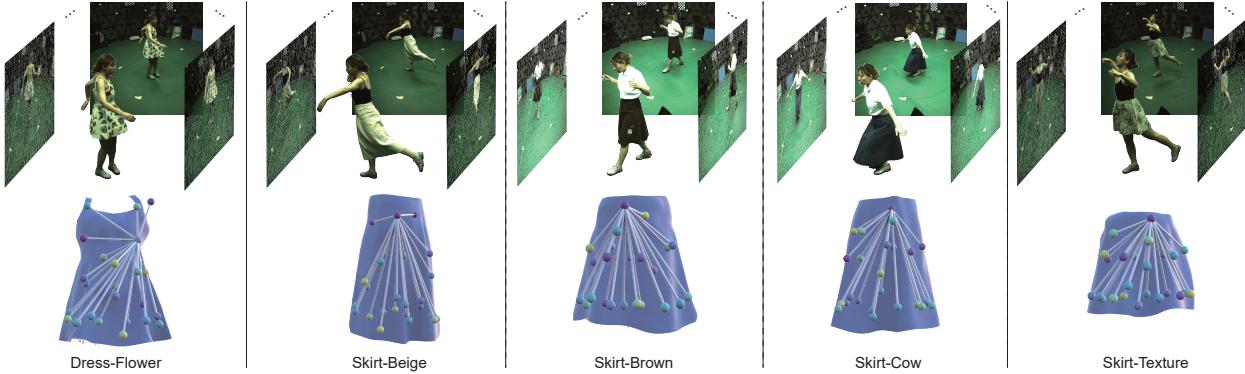
Supplementary Material

Figure 9. Overview of our captured dataset. We show the captured multi-view images for each garment in row 1 and their corresponding rigging model in row 2. Specifically, we captured 5 different loose clothes denoted as "Dress-Flower", "Skirt-Beige", "Skirt-Brown", "Skirt-Cow" and "Skirt-Texture" respectively.

7. Dataset Overview

Fig. 9 presents an overview of our captured dataset, encompassing 5 different loose clothes, namely "Dress-Flower", "Skirt-Beige", "Skirt-Brown", "Skirt-Cow" and "Skirt-Texture". For each garment, we recorded a minimum of two sequences, referred to as "Seq-1" and "Seq-2". Additionally, Fig. 9 shows the garment template mesh and its associated rigging model in the second row.

8. Implementation Details

We provide more implementation details in this section.

Garment T-pose Reconstruction. Our method necessitates a template mesh for each garment to run physics-based simulations. Contrary to previous approaches [13] which relied on costly 3D scanners to obtain such templates, we adopt a more economical approach utilizing recent progress in neural surface reconstruction. We employ a monocular camera to photograph subjects in T-pose and further calibrate the cameras with structure-from-motion techniques [49, 50]. The signed distance function in T-pose is reconstructed using NeuS [59] and the mesh is extracted via marching cubes [30]. To isolate garment parts, we initially segment the full-body mesh using coarse labels from 2D human parsing [23], followed by manual corrections for inaccuracies. The reconstructed mesh is then downsampled to a vertex count between 6K and 9K.

Garment Rigging Model Extraction. To build a rigging model, we first accumulate a variety of body movements for physics simulations. We utilize 6,000 frames of high-quality, self-penetration-free human motion data from AMASS [34], including walking, running, and dancing. This is supplemented with 4,000 frames of our own captured body motion. We conduct the simulation on each garment using Houdini Vellum [51] with default settings except for the fabric's bending stiffness and the simulator's time scale. The bending stiffness is set to 1e-4 for all garments, offering a balance between dynamic richness and avoiding unnatural simulations. The time scale is adjusted to 1.5 for AMASS motions and 0.3 for our captures, based on their respective motion capture FPS. After simulation, we extract a garment rigging model using the open-source tool provided by [21] from the simulated garment mesh sequence. Specifically, we set the iteration count to 50 to ensure convergence. The number of bones B to 25 for each garment empirically.

Garment Dynamics Fitting. For the first frame, we only use silhouette loss and normal loss since there is no optical flow prediction for the first frame. To estimate the garment poses at t-th frame, we use the garment poses from $(t-1)$-th frame for initialization and optimize it using all three losses. Specifically, we set loss weights as $\lambda_M = 2.0, \lambda_N = 0.1, \lambda_{of} = 3.0$.

Pose-driven Rendering Network. For the total training loss function, we set the weight $\lambda = 0.2$ for all sequences. Except for the appearance neural radiance filed $\mathcal{F}_c$ in canonical space, our pose-driven rendering network needs to jointly optimize a motion deformation filed $\mathcal{W}$ conditioned on both body and garment pose. To better solve the motion field, following [62], we initialize $\mathcal{W}$ with the forward
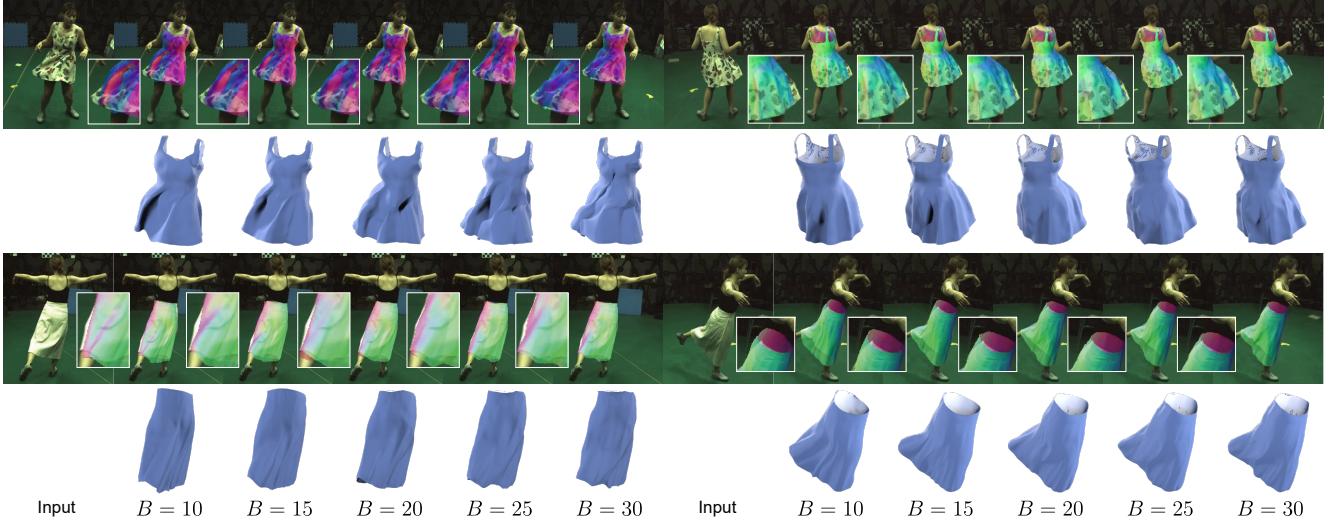
Figure 10. Garment fitting results with different numbers of garment bones B. In the 1st row, we show the original image and the rendered normal images of captured meshes. The fitted meshes are shown in the 2nd row for each case.

skinning weights given by the canonical T-pose and ask the network to learn the residue. For the body part, we also use the combination of ellipsoidal Gaussian around each body bone, like [62]. However, for our garment rigging model, the inside 'virtual' bones are not tied to the garment surface (as shown in Fig. 9). Therefore, we initialize $\mathcal{W}$ with the skinning weights around garment T-pose mesh vertices for the garment part.

9. More Experiment Results

9.1. Garment Dynamics Fitting

Number of Bones. The number of garment bones B serves as a critical hyperparameter in constructing rigging models. For all experiments presented in our main paper, we set B as 25. In this subsection, we evaluate the fitting performance of our method on "Dress-Flower-Seq3" and "Long-Beige-Seq1" using different numbers of garment bones (10, 15, 20, 25, 30). The fitting results are presented in Fig. 10. We only provide qualitative results, as there is no ground-truth 3D garment geometry. The first row displays the original images and the corresponding normal images of the fitted meshes, while the second row shows the reconstructed meshes. As indicated in Fig. 10, models with fewer bones lack the expressiveness to capture garment folds and shapes accurately. Generally, models with more joints yield more precise fitting results. Our approach primarily recovers folds and shapes observable in the images, excluding the finer details of clothing wrinkles. This limitation arises mainly due to two factors. First, a more expressive rigging model capable of detailing wrinkles is required, which is beyond the scope of our current method. The LBS model, derived from skinning decomposition, represents the folds

Figure 11. Garment fitting results under 4 cameras. We show the original image in the 1st row and the rendered normal images of captured meshes in the 2nd row. The fitted meshes are shown in the 3rd row.

and shapes of simulated data but omits local wrinkles, making it less effective for capturing local deformations. Second, our method depends on 2D cues like image normals and optical flows predicted from [55, 67, 68] for capturing temporal consistent 3D folds. While these cues are reliable for large folds and shapes, they are less accurate in depicting local wrinkles.

Number of Views. All experiments in our main paper are conducted, setting the number of views N^v as 8. Here, we further decrease the number of views N^v to 4 and present qualitative results in Fig. 11. As shown in

2

Table 4. Quantitative results of novel view synthesis. Generally, our method reports slightly lower PSNRs than NB (NeuralBody) [42] while outperforming all baselines on LPIPS [73] and DISTS [9]. We color cells that have the best and second best scores. (AN: Anim-Nerf [41], NB: NeuralBody [42], Uv: UV-NeRF [5] and w/o **G**: without using garment pose.)

Garment	Sequence	PSNR ↑					LPIPS ↓ [73]					DISTS ↓ [9]				
		AN	NB	Uv	w/o **G**	ours	AN	NB	Uv	w/o **G**	ours	AN	NB	Uv	w/o **G**	ours
Dress-Flower	Seq1	27.26	29.56	27.31	28.25	28.76	0.0492	0.0318	0.029	0.0231	0.0217	0.1389	0.0928	0.1002	0.0824	0.0783
	Seq2	26.15	28.62	26.75	26.25	27.62	0.0575	0.0382	0.0324	0.03	0.0247	0.1643	0.1047	0.1065	0.1006	0.0851
	Seq3	26.06	28.62	26.14	26.5	27.29	0.0614	0.0393	0.0355	0.0298	0.0263	0.1744	0.1091	0.1154	0.0975	0.0865
	Seq4	25.39	28.73	26.49	27.01	27.61	0.0631	0.0411	0.0346	0.0282	0.0255	0.177	0.1104	0.1094	0.0902	0.0838
Skirt-Beige	Seq1	28.05	30.36	29.22	29.43	29.5	0.05	0.0346	0.0292	0.0264	0.026	0.1363	0.0884	0.0829	0.0820	0.0832
	Seq2	27.38	30.22	28.64	28.5	28.64	0.0583	0.0392	0.0329	0.0284	0.0286	0.1681	0.1031	0.099	0.0978	0.0961
Skirt-Brown	Seq1	25.9	26.26	25.56	25.92	25.54	0.0627	0.0458	0.0362	0.0317	0.0315	0.1596	0.1153	0.1049	0.0924	0.0929
	Seq2	26.29	26.91	25.41	25.85	25.83	0.0588	0.0488	0.0359	0.0331	0.0305	0.1589	0.1321	0.1069	0.0976	0.0952
Skirt-Cow	Seq1	25.78	26.7	25.52	26.06	25.58	0.0653	0.0541	0.0393	0.0356	0.0338	0.1626	0.1315	0.106	0.0963	0.0982
Skirt-Texture	Seq1	28.44	29.84	27.9	28.96	29.17	0.0565	0.044	0.0323	0.0255	0.0244	0.1546	0.1201	0.0935	0.0818	0.0788
	Seq2	28.43	30.17	28.25	28.84	29.11	0.0551	0.0396	0.0307	0.0261	0.0244	0.1615	0.1149	0.0954	0.0844	0.0803

Table 5. Quantitative results of novel pose synthesis. Generally, our method outperforms baselines on PSNR, LPIPS [73], and DISTS [9]. We color cells that have the best and second best scores. (AN: Anim-Nerf [41], NB: NeuralBody [42], Uv: UV-NeRF [5] and w/o **G**: without using garment pose.)

Garment	Sequence	PSNR ↑					LPIPS ↓ [73]					DISTS ↓ [9]				
		AN	NB	Uv	w/o **G**	ours	AN	NB	Uv	w/o **G**	ours	AN	NB	Uv	w/o **G**	ours
Dress-Flower	Seq1	24.15	24.53	23.85	24.17	25.36	0.0609	0.0553	0.0517	0.0426	0.0387	0.1734	0.1488	0.1605	0.1155	0.1103
	Seq2	24.76	25.00	24.56	25.11	26.18	0.0642	0.0534	0.0501	0.0385	0.034	0.1760	0.1354	0.1616	0.1156	0.1088
	Seq3	24.88	24.85	24.3	25.4	26.11	0.0642	0.0518	0.0483	0.037	0.0337	0.1800	0.1448	0.1495	0.1119	0.1030
	Seq4	24.63	25.21	24.08	24.94	25.85	0.0653	0.053	0.0483	0.038	0.0358	0.1851	0.1436	0.1458	0.1097	0.1051
Skirt-Beige	Seq1	25.37	26.56	25.66	26.52	26.58	0.9383	0.9415	0.9377	0.9445	0.8968	0.1609	0.1500	0.1672	0.1127	0.0840
	Seq2	25.9	26.58	25.85	26.92	27.71	0.0604	0.053	0.0479	0.0403	0.0364	0.1733	0.1392	0.1542	0.1221	0.1222
Skirt-Brown	Seq1	24.1	24.71	24.00	24.24	24.38	0.0632	0.0533	0.0469	0.0418	0.0373	0.1642	0.1405	0.1369	0.1154	0.1101
	Seq2	23.64	24.43	23.25	23.95	23.82	0.0701	0.0609	0.053	0.0456	0.0419	0.1778	0.1522	0.1467	0.118	0.1137
Skirt-Texture	Seq1	26.92	27.63	26.52	27.29	27.91	0.0576	0.0501	0.0417	0.0342	0.0314	0.1628	0.1384	0.1272	0.1025	0.0946
	Seq2	26.43	27.9	26.53	27.23	27.46	0.063	0.0489	0.0415	0.0359	0.0351	0.1714	0.1292	0.1106	0.0988	0.0991

Fig. 11, our method still produces reasonable fitting results consistent with 2D observations, indicating the robustness of our method. However, for complex garment geometries like "Dress-Flower", our method can cause unnatural mesh distortions when capturing complex motions like dress whirling.

9.2. Pose-driven Rendering Network

In this subsection, we evaluate the performance of our method for novel view synthesis on "Dress-Flower-Seq3".
Metrics. We use PSNR, LPIPS [73] and DISTS [9] to evaluate the results. PSNR evaluates the difference between predicted and ground-truth images at pixel level while LPIPS [73] measures their distance at feature level and agrees well with human visual perception. Moreover, DISTS [9] measures the structure and texture similarity between the predicted images and the ground truth.
Motion Weights Initialization. We compare different initialization strategies mentioned in Sec. 8. As shown in Tab. 6, using the blend weights prior from garment T-pose mesh vertices instead of the virtual bones will improve the synthesized image qualities.
Number of Bones. All experiments in our main paper are conducted with the number of garment bones B as 25. Here, we show results with different numbers of bones

(10/15/20/25/30) in Tab. 7. In this setting, both fitting and rendering are conducted using the same number of joints. We can observe that $B=25$ achieves the best results in terms of all scores, further validating our choice.

Table 6. Novel view synthesis results of our models trained without and with motion weights initialization using garment T-pose mesh vertices.

	PSNR ↑	LPIPS ↓	DISTS ↓
w/o vertices prior	26.86	0.0272	0.0906
w/ vertices prior	27.29	0.0263	0.0865

Table 7. Novel view synthesis results of our models trained with different numbers of garment bones B.

Number of Bones	PSNR ↑	LPIPS ↓	DISTS ↓
$B=10$	27.03	0.0274	0.0912
$B=15$	26.95	0.0286	0.0935
$B=20$	26.96	0.0279	0.0923
$B=25$	27.29	0.0263	0.0865
$B=30$	27.12	0.0275	0.090

Number of Views. All experiments in our main paper are conducted with the number of views N^v as 8. Here, we

Table 8. Novel view synthesis results of our models trained with different numbers of camera views N^v.

Cameras	PSNR ↑	LPIPS ↓	DISTS ↓
4	27.02	0.0273	0.0882
8	27.29	0.0263	0.0865

show results using 4 training views in Tab. 8 (both fitting and rendering are conducted using 4 cameras). This experiment demonstrates that our method works well on sparse views even with only 4 cameras.

9.3. Comparisons on Per-sequence

In our main paper, we report the average scores of our method in terms of PSNR, LPIPS [73] and DISTS [9] across the entire dataset. Here, we provide qualitative comparisons of our method against baselines for each sequence in both novel view and novel pose synthesis, as detailed in Tab. 4 and Tab. 5 respectively. For brevity, we abbreviate Ani-NeRF [41] as AN, NeuralBody [42] as NB, UV-Volumes [5] as UV in these tables. For novel view synthesis, our method demonstrates marginally lower PSNR values compared to NeuralBody[42], yet surpasses all baselines in LPIPS scores. This suggests that our method yields results agreeing more with human visual perception. Additionally, our method achieves the lowest DISTS scores, implying that our results are more similar to ground-truth images in terms of global structures and local textures. For novel pose synthesis, our method outperforms all baselines in terms of PSNR, LPIPS and DISTS, showing its superior capability in novel pose generalization.