DOI: 10.1111/cgf.15178

ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2024
M. Skouras and H. Wang
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 8

Garment Animation NeRF with Color Editing

Renke Wang, Meng Zhang[†], Jun Li, Jian Yang[†]

PCA Lab,
Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education,
and Jiangsu Key Lab of Image and Video Understanding for Social Security,
School of Computer Science and Engineering, Nanjing University of Science and Technology
{wrk226, mengzephyr, junli, csjyang}@njust.edu.cn

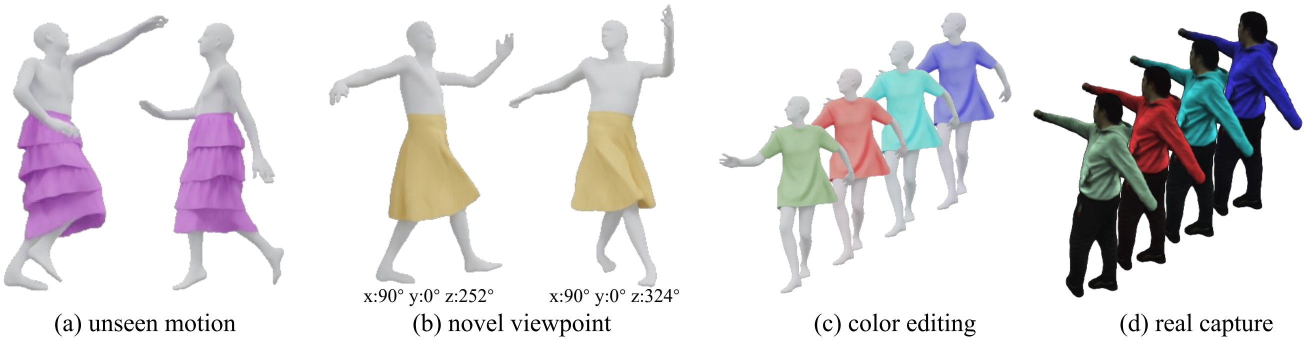


Figure 1: *We introduce a novel* Garment Animation NeRF *that generates character animations directly from body motion sequences, eliminating the need for an explicit garment proxy. Upon training, our network produces garment animations with intricate wrinkle details, ensuring plausible body-and-garment occlusions and maintaining structural consistency across views and frames. We demonstrate the network's generalization capabilities across unseen body motions (a) and camera views (b), while also enabling color editing for garment appearance (c). Notably, our method is applicable to both synthetic (a, b, c) and real-capture (d) garment data.*

Abstract

Generating high-fidelity garment animations through traditional workflows, from modeling to rendering, is both tedious and expensive. These workflows often require repetitive steps in response to updates in character motion, rendering viewpoint changes, or appearance edits. Although recent neural rendering offers an efficient solution for computationally intensive processes, it struggles with rendering complex garment animations containing fine wrinkle details and realistic garment-and-body occlusions, while maintaining structural consistency across frames and dense view rendering. In this paper, we propose a novel approach to directly synthesize garment animations from body motion sequences without the need for an explicit garment proxy. Our approach infers garment dynamic features from body motion, providing a preliminary overview of garment structure. Simultaneously, we capture detailed features from synthesized reference images of the garment's front and back, generated by a pre-trained image model. These features are then used to construct a neural radiance field that renders the garment animation video. Additionally, our technique enables garment recoloring by decomposing its visual elements. We demonstrate the generalizability of our method across unseen body motions and camera views, ensuring detailed structural consistency. Furthermore, we showcase its applicability to color editing on both real and synthetic garment data. Compared to existing neural rendering techniques, our method exhibits qualitative and quantitative improvements in garment dynamics and wrinkle detail modeling. Code is available at `https://github.com/wrk226/GarmentAnimationNeRF`.

CCS Concepts

• ***Computing methodologies*** → *Rendering; Animation; Neural networks;*

[†] Corresponding author

1. INTRODUCTION

High-fidelity, detailed garment animation is crucial for enhancing user engagement across various applications such as games, movies and virtual/augmented reality. Modeling, simulating and rendering dynamic garments with realistic folds and wrinkles necessitates a computationally intense processing pipeline, typically operated on expensive professional setups. Any further update to garment appearance, dynamics or viewing camera positions is a tedious task that requires repeating large portions of the pipeline.

Neural rendering represents a significant breakthrough, enabling to learn neural features for controllable image synthesis, including changes in viewpoint and modeling deformations [TTM*22]. The neural rendering methods in garment context [LHR*21, CWC*23, XFM22, PZX*21], learns integrating clothing simulation and rendering from multi-view image data. Such techniques enable to synthesize the animation of actors wearing target garments. An intriguing alternative involves the use of learning-based method to translate dynamic garment video synthesise from 2D human cues, such as body pose [ASL*19, CGZE19, DLS*19], body-part segmentation [ZWF*19]. However, those methods are restricted to tight clothing and don't have good performance on cases of loose and complex garments under dynamic body motions. Dynamic Neural Garments (DNG) [ZWCM21] employ learnable neural texture for a coarse garment proxy and translate neural descriptor maps to dynamic garment appearance renderings. While demonstrating impressive results with vivid fine wrinkle details, due to a lack of 3D spatial awareness during network inference, DNG struggles to maintain detailed structural consistency across dense views, and requires resolving of the garment-and-body occlusion in the post-processing.

In this work, given a sequence of body motion as input, our work aims at synthesizing dynamic garment animations with fine wrinkle details, without coarse garment proxy predefined [ZWCM21] or explicitly reconstructed from multi-view images [HLX*21, XBS*22, XPC*23], meanwhile, ensuring detailed structural consistency across views. On supervision of multi-view video data, we propose an architecture of neural networks to learn (i) garment dynamic mechanism, (ii) fine wrinkle details, (iii) and structural consistency across view points. Though trained garment-specific, our network has the generalization ability to synthesize target garment animations for a variety of human body shapes and motion sequences that are reasonable close to the training data. Moreover, we achieve an artefact-free color editing of the target garment appearance.

To this end, we first infer a garment dynamic feature map from the geometric and dynamic information of the body, by recording the historical data in the body template texture space. This map provides a preliminary structural overview of the garment's dynamics influenced by the body's movements. Simultaneously, we encode wrinkle detail feature maps from synthesized garment images using a pre-trained image generative model. To minimize wrinkle detail discrepancies between views caused by the generative model, we carefully select synthesized images from the front and back views to generate detail feature maps with minimal overlap. In the final rendering stage, we sample the dynamic and detail features according to the radiance point projection on the body's geometry. Additionally, we calculate body-relative geometric features for the radiance points to enhance the 3D spatial awareness with respect to the moving body. Concatenating those features as input, we construct a neural radiance field (NeRF) [MST*21] to render garment appearance feature images. Furthermore, we utilize a decoder network to decompose the appearance feature image into several components according to palette base colors to achieve neat segmentation of the target garment. By disentangling the light effect from the pixel color and computing semantic layer-based blending weights, we allow the color editing of garment appearance, maintaining the detail structural realistic. Consequently, our method efficiently renders plausible garment animations driven by body motions, maintaining detailed structural consistency across dense views. Additionally, it facilitates the recoloring of garment appearances, enhancing the adaptability and utility of the rendered animations. For example, it allows for a quick preview of a garment in various colors.

We evaluate our algorithm on a variety of real and synthetic garments with varying body motions. We showcase that our method generalizes effectively across different body motions and dense view rendering. Additionally, we edit the color of garment appearance, preserving realistic dynamic garment structure. Our method quantitatively and qualitatively surpasses three benchmark techiques: SANeRF [XFM22], UVVolume [CWC*23], and DNG [ZWCM21]. It delivers convincing garment dynamics, distinctive wrinkle details, accurate body-and-garment occlusions (see Figure 8 and Table 1).

In summary, our key contributions are:

- *body-motion-based NeRF* to synthesize the animation of a complex target garment, only taking a body motion as input, with no need of an explicit coarse garment proxy;
- *generalization* over unseen body motions to render garment animations with fine wrinkle details, maintaining structural consistency across views and frames;
- *palette-based neural rendering* to allow color editing of target garment appearance free-from artefacts.

2. RELATED WORK

2.1. Modeling garment dynamics

The employment of physically based simulations has demonstrated effectiveness in achieving authentic depictions of cloth dynamics [CK05, LLK19, NSO12, NMK*06, TWL*18, YZZ*19]. While yielding highly accurate results, this approach is notably computationally intensive. Although subsequent endeavors have introduced acceleration schemes [LTT*20, WWYW20], the computational efficiency and stability of these methods remain still largely impacted by the complexity of the garment.

To tackle these challenges, data-driven approaches have emerged by involving modeling garment deformations based on pose and body shape [GRH*12, MYR*20, PLPM20, SOC19, LZWL23], typically utilizing linear blend skinning. While effective and computationally efficient, they are constrained by their reliance on skinning, which limits their capacity to represent loose garments like long skirts. To handle this, some approaches [STOC21, LZZ*22] propose diffusing skinning weights from the body to the garment's surroundings to capture loose garment dynamics. Alternatively, Pan et. al. [PMJ*22] introduce virtual bones dedicated to garments to model their dynamics, showing promise in handling loose garments. Unfortunately, these methods often necessitate extensive physical simulations to acquire sufficient 3D data for training. To mitigate this acquirement, physically based constraints as energy losses are leveraged for unsupervised training. For example, Grigorev et. al. [GBH23] employ hierarchical graph neural networks constructed from garment mesh vertices and edges to transfer physical information, while Bertiche et. al. [BME22] directly predict garment deformations based on human motions. These techniques bypass the need for 3D ground truth and have demonstrated impressive performance. However, they face limitations regarding available garment templates and computational resources, and struggle to address self-intersection issues within garments, posing challenges in modeling garments with complex geometries.

In contrast, our model does not necessitate 3D ground truth or garment templates as input, and learns the garment geometry and dynamics directly from multi-view images. This circumvents the manual modeling process and eliminates the requirement for self-intersection detection, significantly augmenting our capability to represent garments with intricate geometries.

2.2. Neural Rendering

In recent years, neural rendering has demonstrated remarkable success in generating novel views of both static scenes [MST*21, WLL*21, KKLD23] and dynamic scenes [PCPMMN21, TTG*21, CJ23, FKMW*23, LKLR23]. As a result, researchers have begun exploring the application of neural rendering techniques to model garment dynamics, aiming to provide an integrated solution for simulation and rendering. A key concept in this endeavor is to learn neural features based on 3D human representations, such as skeleton [ZWCM21, KGE*21, KRKG24, NSLH21, LCY*23, KLF*24] or human body [PZX*21, LHR*21, XFM22, CWC*23, GWL*23, WCS*22], to control rendering output and accurately reflect body motion and corresponding garment dynamics. For example, NeuralBody [PZX*21] utilizes a vertex-based latent code to establish temporal correspondence and employs a spatially sparse convolution network to convert these codes into a radiance volume, producing high-quality results in novel views. Unfortunately, its rendering quality is only guaranteed for motions seen during training. To address this limitation, researchers have proposed various approaches to establish shared appearance carriers for different poses. Neural Actor [LHR*21] utilizes a canonical space combination with inverse skinning to enable unseen pose generalization. Surface-aligned NeRF [XFM22] constructs an implicit field aligned with the mesh surface to capture surface-dependent appearance, while UV Volume [CWC*23] decomposes a dynamic human into 3D UV volumes and a 2D texture map, effectively reconstructing consistent appearance under different poses. However, these methods typically model each frame of motion separately, overlooking the continuity of body motion and encountering difficulties in capturing the dynamics of loose garments.

In contrast to previous methods, we incorporate both velocity and normal extracted from human motion as dynamic features in dynamic modeling. This enables us to achieve dynamic-aware and structure-consistent appearance under unseen motion. Additionally, we employ a 2D detail generator to provide wrinkle details for the target garment, enhancing the richness of detail in our results.

2.3. Image-to-image Translation

The advent of advanced generative models [GPAM*14,HJA20] has facilitated image translation across domains, leading to the proliferation of high-quality translation techniques [CZS17, CUYH20, CLPL23,KZZ*23]. Early successes like Pix2Pix [IZZE17] demonstrated impressive results by aligning pixel data for paired image translation, effectively transferring both style and content. Expanding on this, some studies have utilized continuous driving signals, such as key points, to achieve realistic video generation [NSO12, SLT*19a, SLT*19b]. In the realm of human animation and virtual try-on, some approaches focus on using guides such as semantic maps [ESO18,FLJ*22], skeletons [ASL*19,CGZE19,DLS*19], or dense correspondences [WLZ*18, LPM*19, ZSZS19]. These cues aid in aggregating information across frames and generating coherent representations across different movements. However, these methods often struggle to model loose garments due to challenges in establishing pixel-level correspondences. DNG [ZWCM21] proposed using a coarse garment as a proxy to capture human motion features, enabling realistic garment-driven effects from various viewpoints. Nonetheless, this approach faces difficulties in maintaining detailed structural consistency across different views due to a lack of 3D spatial awareness during inference. Comparatively, our method involves leveraging the detail map generated in image-to-image schemes and aligning it with a pose-aware neural radiance field, resulting in garment dynamics that are not only detailed but also structurally consistent.

2.4. Garment authoring

With the advancement of digital fashion, several algorithms have emerged enabling intuitive garment editing for non-professionals. These include generating 3D models from photos [ZCF*13] or sketches [DPS15, WCPM18], resizing garments for different body shapes [MWJ12], designing textures from wallpapers [WSH19], and crafting various styles of tight-fitting clothes [KZW*15]. Recent developments also facilitate direct adjustments on 3D garments [PDF*22] without simulations. Our research extends these technologies by focusing on garment animation modeling and allowing color editing without artifacts.

3. OVERVIEW

Given a set of multi-view RGB videos captured circle around an motion sequence of a character dressed in a complex or loose target garment, our method is designed to synthesize garment animation in image space when given a new (unseen) human motion sequence and arbitrary view points. We render garment animations with high-quality fine wrinkles, ensuring structural consistency across dense-view rendering. Moreover, our technique allows for color editing of the target garment.

We propose an architecture of *Garment Animation NeRF*, that composes of three components: the *dynamic feature encoder* $\mathcal{E}$, the *detail feature generator* $\mathcal{G}$, and the *rendering network* $\mathcal{M}$. We show the network architecture in Figure 2. Initially, we learn the dynamic features of the garment influenced by the human body motion, by encoding the texture of body motion features using an encoder network $\mathcal{E}$. Simultaneously, to generate the garment dynamic wrinkle features, we utilize a generative model $\mathcal{G}$ to create neural images of garment detail features, taking the front and back views of the body motion as input. In the final stage, we sample the garment dynamic and detail features based on the radiance point projection and compute the 3D body-aware geometric information, which are then concatenated as inputs for the NeRF $\mathcal{M}_{NeRF}$ to render the feature image that depicts the garment appearance. Furthermore, we develop an innovative network $\mathcal{M}_D$, that integrates with $\mathcal{M}_{NeRF}$, for palette-based decomposition, enabling the color editing of the target garment. This is achieved by decomposing the color elements from the appearance features, and then recombining them through a linear combination, thus generating the final target garment image.

4. ALGORITHM

4.1. Dynamic Feature Encoder

Our method first infers garment dynamic structural features from the input of desired 3D body motion. Given a body motion B_t at current time t, we employ UV texture space to capture body geometric and dynamic information, thereby obtaining the body shape normal map $N_t \in \mathbb{R}^{128\times128\times3}$ and the velocity map $V_t \in \mathbb{R}^{128\times128\times3k}$, where k denotes the number of historical frames that dynamic information V_t contains. Utilizing the body information texture $\{N_t, V_t\}$, we employ the dynamic feature encoder $\mathcal{E}$, comprised of a series of 2D convolution layers, to encode the dynamic features of garments. We compute the implicit dynamic feature map $F_t^s \in \mathbb{R}^{128\times128\times8}$ by transforming the body texture features as:

$$F_t^s = \mathcal{E}(N_t, V_t). \tag{1}$$

The dynamic feature map F_t^s of the garment provides a preliminary structural overview of the garment dynamics and the interactions between body and garment. This aids in ensuring garment structural consistency across views and in generating plausible occlusions between the body and the garment. Please note that, unlike methods that map dynamic features to the garment UV space [ZWCM21, HLX*21, XBS*22, XPC*23], we infer garment dynamic features soly from body dynamics, eliminating the need to project features onto any predefined garment proxy space.

4.2. Detail Feature Generator

Inspired by DNG [ZWCM21], we employ a generative model $\mathcal{G}$, to synthesize the detail features of garment wrinkle details, with the desired body motion serving as input. In contrast to DNG, which uses a coarse garment proxy to guide the dynamic neural rendering, our approach directly apply the learnable neural textures to the body geometry and renders neural images Q_t to serve as inputs of the detail feature generator $\mathcal{G}$.

Our detail Feature generator, $\mathcal{G}$, is composed of two parts: the image generator $\mathcal{G}_{img}$ and the detail feature encoder $\mathcal{G}_{en}$. Following the methodology of DNG, we pre-train $\mathcal{G}_{img}$ with learnable neural texture images that capture body dynamics, to predict reference images $\hat{I}_t$ of garment dynamics, sized at 512×512. The training is supervised by ground truth images of the garment. The pre-trained image-based generative model $\mathcal{G}_{img}$ provides essential hints for rendering garment appearance details.

Taking body neural texture images of the body $Q_t(c)$ at a specific camera pose c as input, we use the image generator $\mathcal{G}_{img}$ to predict the reference image $\hat{I}_t(c)$. Then we use the detail feature encoder $\mathcal{G}_{en}$ to generate the detail feature maps $F_t^d(c) \in \mathbb{R}^{512\times512\times128}$ as:

$$F_t^d(c) = \mathcal{G}_{en}[\hat{I}_t(c)] = \mathcal{G}_{en}[\mathcal{G}_{img}[Q_t(c)]]. \tag{2}$$

To mitigate the negative impacts of detail inconsistencies between views, as observed in [ZWCM21], we employ the image generator $\mathcal{G}_{img}$ to render only the front and back views to compute garment detail feature maps, $F_t^d(c^{fr})$ and $F_t^d(c^{ba})$. This ensures that the rendered details features maps are free from detail structural conflicts. Please note that, as trained with multi-view videos captured circle around the character wearing the target garment, the image generator $\mathcal{G}_{img}$ has the generalization ability to render the garment animation at front and back viewpoints, even if they are located outside the training camera viewpoints.

4.3. Rendering Network

Our rendering network $\mathcal{M}$ is mainly composed of a NeRF network $\mathcal{M}_{NeRF}$ to render garment appearance feature image, and a network $\mathcal{M}_D$ for palette-based decomposition to allow color editing of the target garment.

Appearance feature image. Our NeRF network $\mathcal{M}_{NeRF}$ constructs a temporal-and-spatial aware dynamic implicit field, adaptive to body dynamics. It aims to learn three key factors: (i) the garment's dynamic structure driven by body motion;(ii)wrinkle details enhancement ensuring temporal coherence; and(iii) occlusions between garment layers or between the garment and the body. To achieve, we first construct 3D garment dynamic feature distribution with features sampled from the acquired dynamic structural feature map F_t^s. Next, we project generated detail features back to 3D space to concatenate them with the garment features and body-aware spatial information. Finally, while constructing the 3D dynamic garment appearance field, we compute a density field around the moving body to indicate the occlusion between the garment and the body, or within the garment's structural layers.

With a 3D point position x sampled along the camera ray towards a specified image pixel, we begin by projecting x to the posed body geometry B_t to obtain the projection point b_t. We then calculate the distance $h = |x - b_t|$ between x and b_t. To enable pose-invariant body-relative sampling, we transform b_t back to the canonical pose, yielding the position b_o on the canonical body shape B_o. This allows us to represent the body-aware geometric information as $x_t^b := [b_o, h]$. Utilizing the corresponding body UV space coordinate of b_o, we sample an implicit garment dynamic feature f_t^s from the acquired dynamic structural feature map F_t^s. Simultaneously, we sample the detail feature f_t^d from the detail feature image rendering, either at the front $F_t^d(c^{fr})$ or back view $F_t^d(c^{fr})$, depending on whether the projection point b_t is visible at the camera pose c^{fr} or c^{ba}.

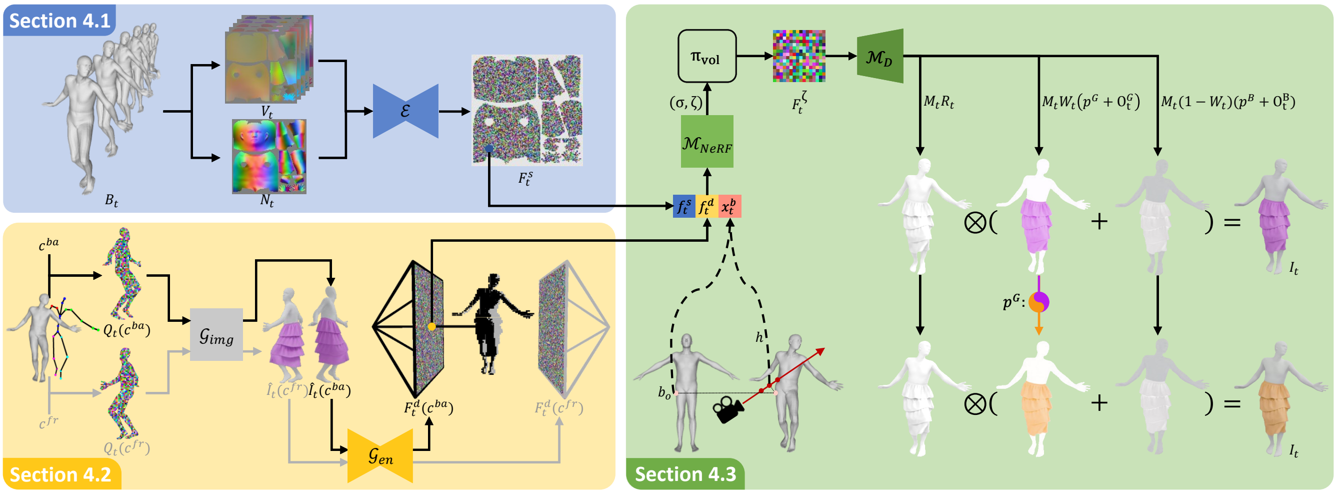


Figure 2: ***The architecture of*** **Garment Animation NeRF.** *Given a sequence of character's body motion, we construct a neural radiance field to render animation of the character dressed in the target garment. We first employ a dynamic feature encoder $\mathcal{E}$ to infer garment dynamic feature map F_t^s from the information textures V_t and N_t of the body motion B_t. Simultaneously, taking body neural texture images Q_t of the body at front c^{fr} and back views c^{ba}, we use a pre-trained image generator $\mathcal{G}_{img}$ to predict reference images $\hat{I}_t(c^{fr})$ and $\hat{I}_t(c^{ba})$. Subsequently, we use the detail feature encoder $\mathcal{G}_{en}$ to generate the detail feature maps $F_t^d(c^{fr})$ and $F_t^d(c^{ba})$. Then, we obtain body-aware geometric information x_t^b by calculating the distance h between sampling points and the body surface, and finding b_o on the canonical body shape. Finally, we utilize a NeRF network $\mathcal{M}_{NeRF}$ to render garment appearance feature image F_t^ζ. To enable color editing, we introduce a network $\mathcal{M}_D$ to decompose the garment appearance into a front mask M_t, a color offset map O_t, a radiance map R_t and a blending weight map W_t. By linearly recombining those visual elements, we synthesize the final frame image I_t. Except for the generator $\mathcal{G}_{img}$, we jointly train the networks of $\mathcal{E}$, $\mathcal{G}_{en}$, $\mathcal{M}_{NeRF}$ and $\mathcal{M}_D$ in an end-to-end manner.*

Given inputs of dynamic feature sampling f_t^s, detail feature sampling f_t^d, and body-aware geometric information x_t^b, our NeRF network, $\mathcal{M}_{NeRF}$, computes the density $\sigma \in \mathbb{R}^1$ and garment appearance feature $\zeta \in \mathbb{R}^{128}$ at the 3D point position x. This computation is expressed as follows:

$$\sigma, \zeta = \mathcal{M}_{NeRF}[f_t^s, f_t^d, x_t^b], \tag{3}$$

where $\mathcal{M}_{NeRF}$ consists of multiple layers of perceptrons. Subsequently, we employ volume rendering π_{vol} to generate the garment appearance feature image $F_t^\zeta \in \mathbb{R}^{128\times128\times128}$ as:

$$F_t^\zeta = \pi_{vol}(\zeta, \sigma). \tag{4}$$

The volume rendering π_{vol} adheres to the integration methodology detailed in [MST*20], with the modification of substituting the color element with the appearance feature ζ.

Palette-based decomposition. Inspired by [KLB*23], we propose a 2D convolution neural network $\mathcal{M}_D$ to decompose the appearance feature image F_t^ζ into multiple visual elements. These elements are then linearly blended using a predicted weight map to generate the final target garment image.

To extract the palette-based colors of interest, we statistically compute the mean colors, μ^G and μ^B, for both the regions of target garment G and body B. We initialize the palette base color vector $p := [p^G, p^B]$ with $p^* = [\mu^G, \mu^B]$, that $p \in \mathbb{R}^6$. With the acquired appearance feature map F_t^ζ as input, the network $\mathcal{M}_D$ up-samples the feature image from the size of 128 to 512, and meanwhile, decouples multiple visual element maps, specifically, the color offset map O_t, radiance map R_t, blending weight map W_t. This computation is expressed as follows:

$$O_t, R_t, W_t, M_t = \mathcal{M}_D[F_t^\zeta]. \tag{5}$$

To focus on rendering the moving body dressed with the target garment, $\mathcal{M}_D$ also predicts a mask M_t of the region of both garment and body.

For each image pixel i in the decoupled map, its color offset element, $o_t := [o_t^G, o_t^B], o_t \in \mathbb{R}^6$ in the map O_t, captures its offset with respect to the palette-based color vector p. The radiance element, $r_t \in \mathbb{R}^3$ in the map R_t, represents the intensity of light, while its weight element, $w_t \in \mathbb{R}^1$ in the map W_t, determines the likelihood that the pixel i belongs to the garment region. Additionally, $m_t \in \mathbb{R}^1$ in the mask M_t, indicates whether pixel i falls within the rendering region of both the garment and body. Consequently, the color $c(i)$ of pixel i is computed as:

$$c(i) = m_t \otimes r_t \otimes [w_t \otimes (p^G + o_t^G) + (1 - w_t) \otimes (p^B + o_t^B)], \tag{6}$$

where $\otimes$ denotes element-wise vector multiplication. This formulation allows us to modify the color of the target garment by adjusting the palette-based vector p^G to any desired color $\hat{p}^G$, while keeping the other visual elements unchanged.

4.4. Loss Function

Except for the generator $\mathcal{G}_{img}$, which is pre-trained following the instruction in [ZWCM21] supervised by ground truth images as discussed in Section 4.2, we jointly train the dynamic feature encoder $\mathcal{E}$, the detail feature encoder $\mathcal{G}_{en}$, and the rendering blocks of $\mathcal{M}_{NeRF}$ and $\mathcal{M}_D$, in an end-to-end manner.

To ensure the image rendering quality, we employ L1 loss with respect to the colors and masks in the region of interest. We define a loss function L_{img} as follows:

$$L_{img} = \|I_t - I_t^*\|_1 + \|M_t - M_t^*\|_1. \tag{7}$$

where, I_t denotes an image synthesized by our network with I_t^* as its ground truth, and M_t denotes the predicted mask with M_t^* as its ground truth. Furthermore, we consider the similarity of multi-layer features $VGG^i[I_t]$ of the pre-trained network VGG and represent the loss function as:

$$L_{vgg} = \sum_i \|VGG^i[I_t] - VGG^i[I_t^*]\|_1. \tag{8}$$

For color editing of the target garment, we adapt the sparsity loss L_{sp} and the color offset loss L_{off} to achieve correct segmentation of the target garment. The sparsity loss L_{sp} is applied on the predicted blending weight map W_t, and defined as:

$$L_{sp} = \|\frac{1}{W_t^2 + (1-W_t)^2} - 1\|_1. \tag{9}$$

And the offset loss L_{off} is applied on the predicted color offset map O_t, and defined as:

$$L_{off} = \|O_t\|_2^2. \tag{10}$$

Following [KLB*23], the sparsity loss L_{sp} and the offset loss L_{off} act as adversarial roles to achieve a neat segmentation of the target garment region, with the optimization of the palette base color vector p by applying the loss L_p defined as:

$$L_p = \|p - p^*\|_2^2, \tag{11}$$

where p^* is the initialization of the base vector introduced in Section 4.3. The final loss function we use to train our network is a weighted sum of the terms:

$$L = \lambda_1 L_{img} + \lambda_2 L_{vgg} + \lambda_3 L_{sp} + \lambda_4 L_{off} + \lambda_5 L_p, \tag{12}$$

where we set $\lambda_1 = 1, \lambda_2 = 0.1, \lambda_3 = 0.0002, \lambda_4 = 0.03, \lambda_5 = 0.001$ in our experiments. λ_1 and λ_2 were chosen to balance the weights of image and VGG features. λ_3, λ_4. λ_5 were set following [KLB*23].

5. RESULTS AND EXPERIMENTS

In this section, we demonstrate the effectiveness of our approach in various scenarios. We also compare the quality and quantity results of our method against other benchmark solutions. We will first show how our method generalizes to novel camera views and unseen motions. Additionally, we will present the color editing capabilities of our method.

5.1. Data Generation

To train our model, we establish a synthetic dataset. We use SMPL [LMR*15] model as our body template, and then obtain a motion sequence of 800 frames from Mixamo (https://www.mixamo.com/) for training. Next, we design three sets of garment (t-shirt, skirt, and multi-layer skirt) using Marvelous Designer, and perform physical simulations of clothing with the body motion sequence to obtain the ground truth garment dynamics. Subsequently, we select 16 fixed camera positions on a circle around the body shape to render videos of the moving body dressed in our target garments. Under each view, we generate the ground truth animation I_t^* and corresponding ground truth front mask M_t^*.

5.2. Implementation details

Architecture. Our dynamic feature encoder $\mathcal{E}$ takes the concatenation of V_t and N_t as input. It initially employs a convolution layer to map the recorded human body information from $\mathbb{R}^{128\times128\times3(k+1)}$ to $\mathbb{R}^{128\times128\times32}$, where we set $k = 2$. Then it's followed by four convolutional layers that gradually downsample the resolution of the feature map from 128×128 to 8×8 and increase the feature dimensions to 64, 128, 256, 512, respectively. Subsequently, we utilize four convolution upsampling layers to gradually upsample the resolution of feature map back to 128×128 and decrease the feature dimensions to 256, 128, 64, 32. At each stage, the output is concatenated with the corresponding downsampled result in a residual-like connection before being passed through the convolution layer. Thus, the encoder network $\mathcal{E}$ computes the dynamic structural feature map F_t^s sized at $128 \times 128 \times 8$. We apply instance normalization and leakyReLU for all convolution layers except the final one, which uses a tanh activation function.

The detail feature encoder $\mathcal{G}_{en}$ employs a similar autoencoder architecture with residual connections as $\mathcal{E}$. Taking a reference image $I_t(c^{fr})$ or $I_t(c^{ba})$ as input, it first downsamples the resolution from 512×512 to 32×32 and increases the feature dimension from 3 to 512. Then, it upsamples the feature map gradually back to resolution of 512×512 with 4 convolution layers, and decreases the feature dimension to 256, 128, 64, 32, respectively. An additional convolution layer is used to produce the final reference feature map F_t^d with size of $512 \times 512 \times 8$.

$\mathcal{M}_{NeRF}$ comprises six linear layers, each with a latent dimension of 256 and ReLU activation. It takes the concatenation of x_t^b, f_t^d, f_t^s as input with shape $\mathbb{R}^{4+8+8}$. We employ 2 additional linear layers to respectively project the output into $\sigma \in \mathbb{R}^{\mathbf{1}}$ and $\zeta \in \mathbb{R}^{\mathbf{128}}$. We compute the appearance feature map F_t^ζ with accumulation processing along pixel rays, following the methodology of volume rendering.

With the appearance feature map $F_t^\zeta \in \mathbb{R}^{128\times128\times128}$ as input, the network $\mathcal{M}_D$ upsamples the resolution from 128 to 512 with 2 convolution layers, and encodes the feature dimension from 128 to 12. Finally, we split the output feature map into a color offset map $O_t \in \mathbb{R}^{512\times512\times6}$, a radiance map $R_t \in \mathbb{R}^{512\times512\times3}$, a blending weight map $W_t \in \mathbb{R}^{512\times512\times2}$ and a front mask $M_t \in \mathbb{R}^{512\times512\times1}$.

Training. We start by training the image generator $\mathcal{G}_{img}$ according to the approach outlined in DNG [ZWCM21], which is subsequently frozen to facilitate the training of the other modules. Except for $\mathcal{G}_{img}$, we jointly train the networks of $\mathcal{E}$, $\mathcal{G}_{en}$, $\mathcal{M}_{NeRF}$ and $\mathcal{M}_D$ in an end-to-end manner. We set the training batch size to be 1 and utilize an Adam optimizer with a learning rate exponentially decaying from 5×10^{-4} to 5×10^{-5}. With a single RTX 3090 GPU, our network converges after 200k iterations, taking around 32 hours for training.

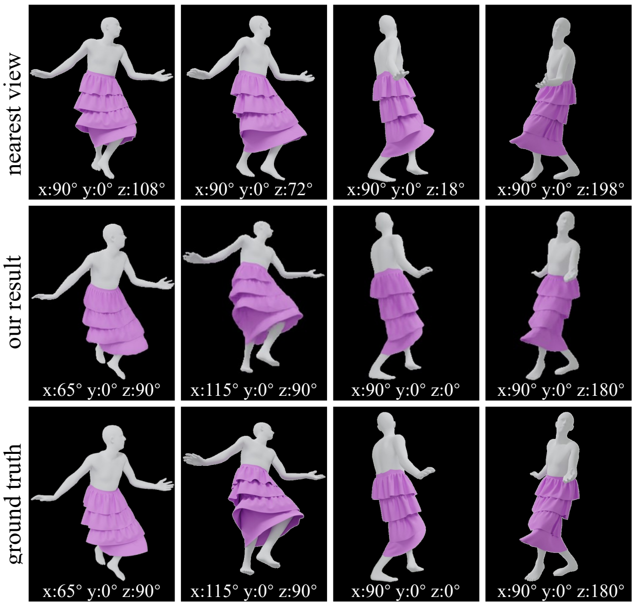


Figure 3: ***Unseen view.*** *Our model can generate a structural consistent garment appearance from arbitrary viewpoint.*

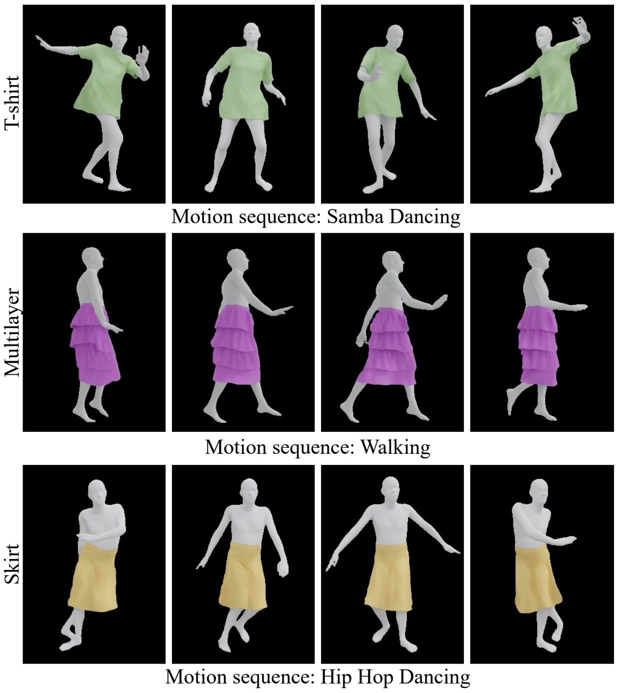

T-shirt
Motion sequence: Samba Dancing
Multilayer
Motion sequence: Walking
Skirt
Motion sequence: Hip Hop Dancing

Figure 4: ***Unseen motion.*** *We test our model on various garments using several motion sequences that were not seen in the training process.*

5.3. Results and evaluation

We now evaluate the generalization ability of our method. Here we show sampled frames and camera views. To better evaluate the visual quality of our garment animation rendering, please refer to the supplementary video.

Unseen views. For the motion sequence seen during training, we demonstrate the generalization of our model across unseen camera view points. In Figure 3, for each test view, we show the training samples with the nearest viewpoint and the ground truth rendering. As can be seen in the supplemental video and Table 1, compared to the nearest training samples, our method generates garment animations in higher performance with respect to perceptual metrics.

Unseen motion. Next, we evaluate our model's performance on motion sequences not seen during training. As illustrated in Figure 4, we conducted tests on the t-shirt, multilayer skirt, and long skirt across different motion sequences, including samba dancing, walking, and hip hop dancing. The results demonstrate that our model exhibits good generalization capabilities to unseen motion sequences. It realistically simulates dynamic effects while preserving garments' topological structure and local folds.

Unseen body shapes. We evaluate the generalization ability to unseen body shapes. DNG [ZWCM21] requires fine-tuning to fit unseen body shapes, as it learns a mapping from the neural texture of the body-independent coarse garment proxy to the target garment animation. In contrast, our method can synthesize garment animations for various body shapes without network fine-tuning, even trained with a fixed body shape. This is because our method synthesizes garment animations based on the distribution of dynamic features f_t^s, detail features f_t^d and body-aware information x_t^b, which are adaptive and closely related to the body shape. Figure 5 demonstrates rendering results of three target garment types tested on two body shapes (round and thin) different from the original body shape of training data.

Real-captured data. We demonstrate that our method is applicable to real-captured garment data. We utilized the multi-view rgb sequence provided by the ZJUMoCap database [PZX*21], along with the fitted SMPL body mesh for each frame. We train the network using 16 viewpoints and 700 frames from the video sequences, while the remaining frames were used to evaluate the model's generalization capability to the unseen motion. As illustrated in Figure 6, our method is capable of generating plausible garment details and wrinkle effects across the unseen body motion.

Color editing. we demonstrate the capability of color editing of our method. As discussed in Section 4.3, the palette-based decomposition of visual elements allows for easy manipulation of colors through adjusting the palette base color vector p. Figure 7 shows the recolored result on both synthetic and real capture data. Compared to the rendering result in original color, the recolored results preserve the detailed folds and wrinkles without introducing artefacts.

Computational performance. Once the training is complete, with our un-optimized python code, our network takes 174 ms per frame in total, to generate the animation of the character dressed in the target garment, including 130 ms to compute the information map $\{N_t, V_t\}$ of the moving body, 5 ms to generate reference images by running the image generative model $\mathcal{G}_{img}$, 39 ms to run the other modules of network. We conduct our experiments on a PC equipped with an AMD 5950X CPU, 64GB of memory, and an NVIDIA GeForce RTX 3090 graphics card.

seen motion												
	t-shirt				skirt				multilayer			
method	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓
SA-NeRF	23.121	0.933	0.136	0.716	21.510	0.915	0.166	0.877	22.085	0.912	0.147	0.766
UV-Volumes	21.520	0.923	0.079	0.835	19.972	0.909	0.113	1.068	20.111	0.896	0.103	0.989
DNG	23.312	0.929	0.091	0.678	23.206	0.927	0.096	0.795	23.150	0.916	0.081	0.703
nearest view	14.976	0.816	0.198	1.305	15.028	0.815	0.194	1.438	15.132	0.806	0.181	1.238
ours	**27.737**	**0.974**	**0.044**	**0.520**	**26.446**	**0.966**	**0.055**	**0.681**	**26.677**	**0.960**	**0.050**	**0.578**
unseen motion												
	t-shirt				skirt				multilayer			
method	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓
SA-NeRF	19.485	0.883	0.197	0.839	17.750	0.862	0.235	1.106	17.967	0.842	0.217	1.022
UV-Volumes	18.822	0.877	0.127	1.073	17.568	0.862	0.158	1.341	17.401	0.837	0.151	1.323
DNG	20.405	0.893	0.140	0.771	19.500	0.883	0.152	0.967	19.562	0.860	0.133	0.893
ours	**22.594**	**0.939**	**0.086**	**0.656**	**20.994**	**0.927**	**0.101**	**0.875**	**21.175**	**0.909**	**0.093**	**0.799**

Table 1: *To evaluate the generalization ability of our method and to compare with baseline methods, we show peak signal-to-noise ratio (PSNR) to qualify the image reconstruction quality of the synthesized images; structural similarity (SSIM) and perceptual similarity (LPIPS) between different methods and the ground truth image; and tOF [CXM*20] to measure the temporal consistency over time.*

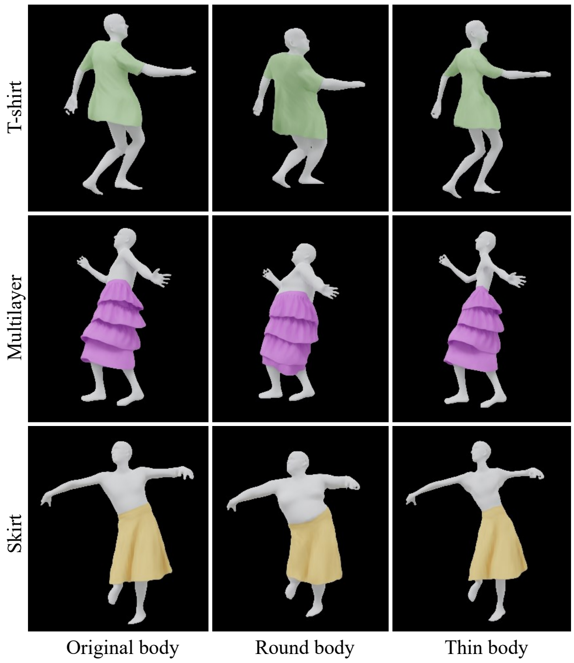


Figure 5: ***Unseen body shapes.*** *Without fine-tuning, our method can synthesize target garment animation that fit with the body shapes (round and thin) different from the original training data body shape.*

5.4. Baseline Comparisons

We compare our method against various baselines, including: (a) Surface Aligned Neural Radiance Fields (SA-NeRF) [XFM22], which constructs neural radiance fields based on relative positions to the human body surface; (b) the method of UV-Volumes [CWC*23], which maps synthesized 2D textures to voxel latent spaces through predicted dense surface correspondences; and (c) Dynamic Neural Garment (DNG) [ZWCM21], which to synthesize the dynamic garment appearance based on learnable neural texture of a garment proxy.

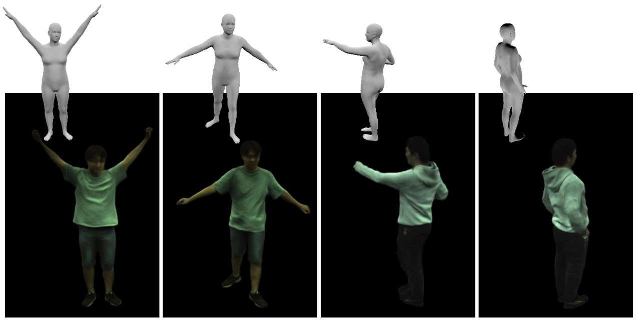
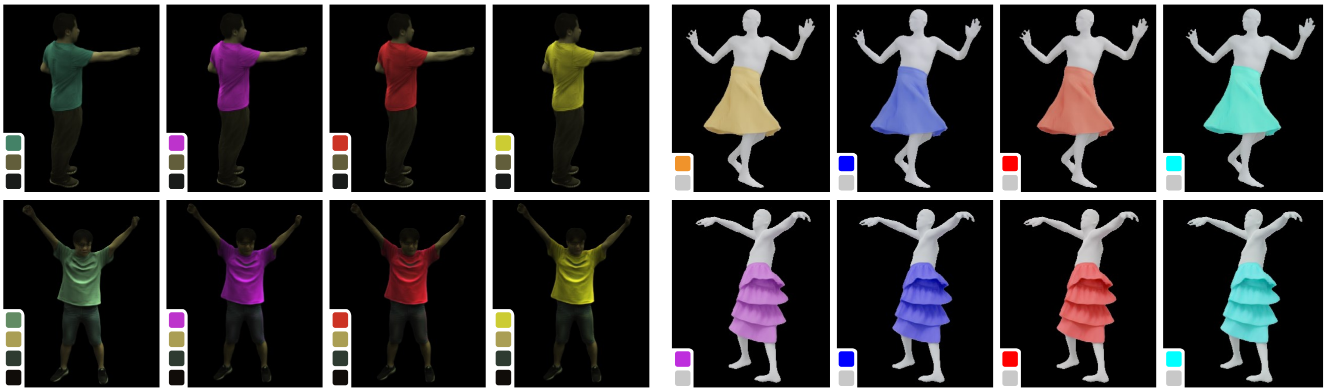


Figure 6: ***Real-captured data.*** *Our model can be trained with real-captured data and generalize to the body motion unseen during the training.*

We train the models of the baseline methods with the same training dataset as ours. The results in Figure 8 demonstrate that although SA-NeRF and UV-Volumes have good performance in handling with the body-and-garment occlusions, they result in garment animation rendering with blurred wrinkle details and incorrect garment shape. Meanwhile, DNG yields visually appealing result but often produces incorrect occlusion relationships in areas like the hands and shoulders, and it can not ensure consistency of the detail structure in dense view rendering that we show the comparison in the supplemental video. In comparison to those baseline methods, our method have better performance in terms of detail richness, contour accuracy, structural consistency and occlusion correctness. In Table 1, we conduct a quantitative comparison demonstrating that our method surpasses the comparative methods in both unseen camera viewpoints and body motions, and achieves higher quality of rendering results with better temporal consistency and more similarity to the ground truth animation.

5.5. Ablation Study

Different training features. We evaluate the effects of detail feature f_t^d and dynamic structural feature f_t^s on our model and present the results in Figure 9 and Table 2. Our results show that models in-



Original Recolored #1 Recolored #2 Recolored #3 Original Recolored #1 Recolored #2 Recolored #3

Figure 7: ***Color editing.*** *By adjusting the palette-based vectors, we can modify the color of the target garment appearance, while maintaining the fabric's fold details.*

used features		seen motion				unseen motion			
detail feature	dynamic feature	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓
		22.160	0.925	0.112	0.765	20.590	0.901	0.130	0.844
✓		26.159	0.956	0.056	0.621	20.689	0.904	0.104	0.878
	✓	26.238	0.957	0.056	**0.574**	21.093	0.908	0.097	**0.764**
✓	✓	**26.677**	**0.960**	**0.050**	0.578	**21.175**	**0.909**	**0.093**	0.799

Table 2: *We present the quantitative evaluation of an ablation study for different input features. The results demonstrate that both detail feature and dynamic structural feature contribute to the improvement of our model's accuracy.*

	seen motion				unseen motion			
# views	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓
2	22.031	0.924	0.097	0.781	19.031	0.882	0.143	0.928
4	26.717	0.969	0.049	0.561	22.399	0.938	0.088	0.689
8	27.340	0.972	0.046	0.545	22.456	0.938	0.088	0.672
16	**27.737**	**0.974**	**0.044**	**0.520**	**22.594**	**0.939**	**0.086**	**0.656**

Table 3: *We report PSNR, SSIM, LPIPS to study the effect of number of camera views during training on the quality of the synthesized image results. We show that training with a greater number of camera views enhances the generalization of our method across unseen camera views and body motions.*

corporating both detail and dynamic structural features outperforms those that include only the geometry feature (x_t^b), or the geometry feature combined with just one of the detail feature (f_t^d) and dynamic structure feature (f_t^s). We observe that integrating either the detail or dynamic feature alone can enhance the rendering quality, improving garment details or contour, respectively.

Different number of training views. When training our Garment Animation NeRF, we set 16 camera poses to generate video of training data. In Figure 10 and Table 3, we demonstrate the results of our model trained with varying numbers of views. Our results indicate that training with a greater number of camera views enhances the generalization of our method across unseen views and unseen body motion. Remarkably, our method can produce plausible rendering results when trained with animation data from only 4 camera views. However, because the primary focus of this paper is not on minimizing the number of views, we chose to train with 16 views in the remainder of our experiments to ensure the highest visual quality possible.

Different number of historical frames. We quantitatively evaluate the impact of using different numbers of historical frames k in the velocity map V_t in Table 4. When $k = 0$ and only normal N_t contributes to generate the dynamic feature F_t^s with $V_t = 0$, our method produces plausible rendering results (with SSIM values above 0.9). The quality of rendering animation for seen motion sequences generally improves as more dynamic information introduced with increasing k. However, for unseen motion sequences, the performance decreases when $k > 2$. We speculate that while historical dynamic information aids in synthesizing garment animation, excessive historical information leads to over-fitting to the training motion sequences, weakening generalization to unseen body motions. When $k = 2$, our method demonstrates a significant generalization to unseen body motions, applicable to all garment types of t-shirt, skirt and multi-layered dress.

6. Limitations and future work

This work presents a novel framework for synthesizing garment animations, especially for complex and loose garments. We directly infer garment dynamics from the movement information of the body, eliminating the need for an explicit garment proxy. However, our work encounters limitations when synthesizing garment dynamics for unseen body motions distributed far away from those

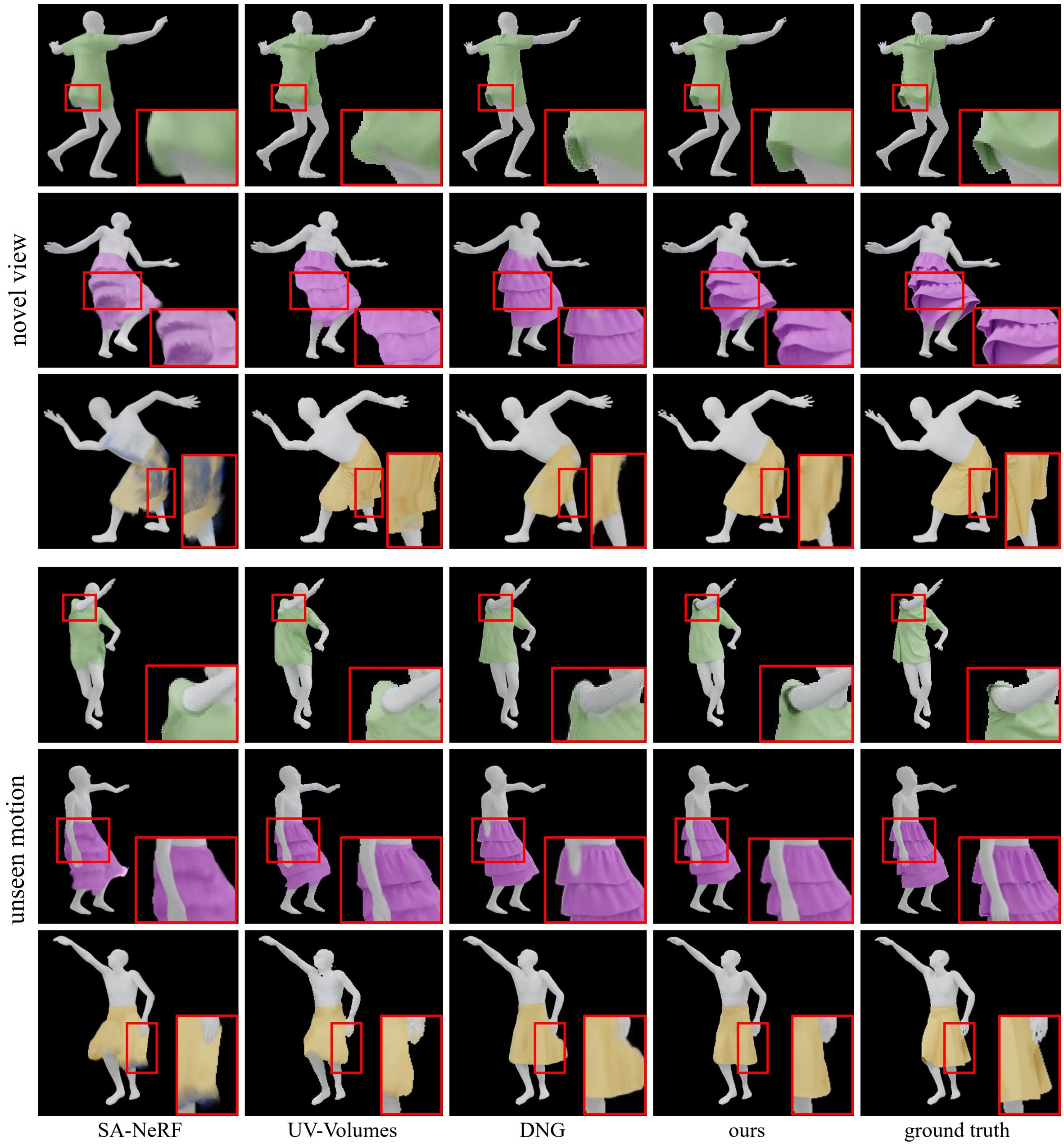


Figure 8: ***Comparisons.*** *We compare our method to SA-NeRF [XFM22], UV-Volumes [CWC*23], as well as DNG [ZWCM21], on cases of both unseen camera views and body motions. Our approach outperforms these baseline methods.*

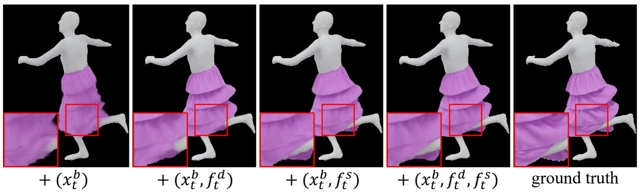


Figure 9: *The effect of different input features.*

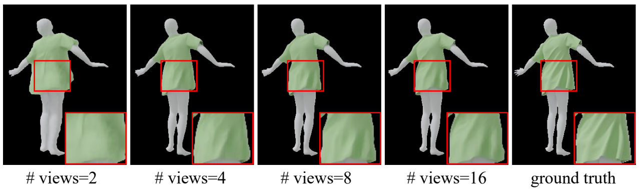


Figure 10: *The effect of number of sampled views during training.*

	seen motion											
	t-shirt				skirt				multilayer			
k	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓
0	27.451	0.973	0.046	0.530	26.458	0.966	0.058	**0.680**	26.380	0.959	0.051	0.584
1	27.522	0.973	0.045	0.541	26.555	0.966	0.058	0.691	26.493	0.959	0.053	0.588
2	**27.737**	**0.974**	**0.044**	**0.520**	26.446	0.966	**0.055**	0.681	26.677	0.960	0.050	0.578
5	27.725	**0.974**	**0.044**	0.525	26.587	0.966	0.057	0.688	26.686	0.960	**0.049**	0.583
10	27.592	**0.974**	0.045	0.537	26.533	0.966	0.057	**0.680**	**26.704**	**0.961**	**0.049**	0.577
20	26.636	0.967	0.055	0.683	**26.636**	**0.967**	**0.055**	0.683	26.593	0.960	**0.049**	**0.570**
	unseen motion											
	t-shirt				skirt				multilayer			
k	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓
0	22.423	**0.939**	0.087	0.661	20.990	0.927	0.104	0.877	21.053	0.908	0.095	0.797
1	22.518	**0.939**	0.087	0.664	**21.192**	**0.928**	**0.101**	0.876	21.135	**0.909**	0.095	0.799
2	**22.594**	**0.939**	**0.086**	0.656	20.994	0.927	**0.101**	0.875	**21.175**	**0.909**	**0.093**	0.799
5	22.499	**0.939**	**0.086**	0.663	21.081	**0.928**	**0.101**	0.877	21.049	0.908	0.095	0.813
10	22.474	0.938	0.088	0.656	21.059	0.927	0.102	**0.864**	21.168	**0.909**	**0.093**	0.799
20	22.528	**0.939**	**0.086**	**0.653**	21.054	0.927	0.104	0.865	21.104	0.908	0.094	**0.796**

Table 4: *We quantitatively evaluate the impact of different historical frame numbers k. When $k = 2$, our method demonstrates a significant generalization ability to unseen body motions, applicable to garment types of t-shirt, skirt and multi-layered dress.*

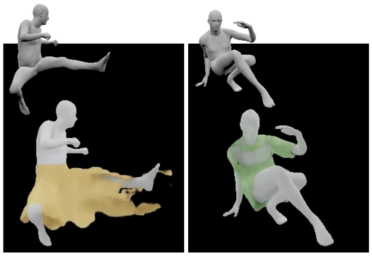


Figure 11: ***Failure cases.*** *Our work falls short when synthesizing garment animation for unseen body motions distributed far away from the training data.*

in the training set, as shown in Figure 11. To address this issue, expanding training dataset is a viable option. Additionally, a more intriguing direction might be to implicitly learn physically-realizable features integrated within the NeRF architecture.

Regarding the capture of wrinkle details, we employ a generative image model to create reference appearance features. Given the rapid advancement in generative AI [CRC*20, MBRS*21, PJBM22] for producing realistic images and videos, we plan to further explore text-driven garment appearance editing in the future work. This exploration will involve generating detailed features tailored with pre-trained text-driven models, such as DALL-E [RPG*21, RDN*22] and Imagen [SCS*22].

Currently, our work supports only color editing of the target garment by decomposing the visual elements of appearance features. Moving forward, we aim to focus more on disentangling the latent features of appearance, differentiating between garment style, texture patterns, materials. This will allow for more flexible and adaptive editing of garment appearances. Moreover, accurately locating a desired 3D position within the implicit field of the dynamic garment constructed by NeRF is challenging. An promising direction for future work is to enable partial color editing of garment animation by transforming the implicit field into an explicitly accessible structural representation [YBZ*22].

Furthermore, we believe the computational time efficiency of our method can be significantly improved in the future work. The most time-consuming steps are the information map recording (130ms) and feature lookup during the NeRF rendering process (30ms). We will try to use graph convolution network [PFSGB20] to directly encode features based on the mesh topology of the body template, instead of running CNN on the 2D feature map. And meanwhile, we intend to incorporate data structures like octrees and hashtables [GKJ*21, YLT*21] into our garment animation NeRF, to more quickly locate relevant features during the rendering inference. It will be promising to enable real-time rendering of garment animation.

7. Acknowledgments

The authors would like to thank the reviewers for their constructive comments; Kun Wang and Xinxin Zou for their valuable suggestions. This work was partially supported by the National Science Fund of China, Grant Nos. 62072242 and 62361166670.

References

[ASL*19] Aberman K., Shi M., Liao J., Lischinski D., Chen B., Cohen-Or D.: Deep video-based performance cloning. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 219–233. 2, 3

[BME22] Bertiche H., Madadi M., Escalera S.: Neural cloth simulation. *TOG 41*, 6 (2022), 1–14. 3

[CGZE19] Chan C., Ginosar S., Zhou T., Efros A. A.: Everybody dance now. In *ICCV* (2019), pp. 5933–5942. 2, 3

[CJ23] Cao A., Johnson J.: Hexplane: A fast representation for dynamic scenes. In *CVPR* (2023), pp. 130–141. 3

[CK05] Choi K.-J., Ko H.-S.: Research problems in clothing simulation. *Computer-aided design 37*, 6 (2005), 585–592. 2

[CLPL23] Cheng B., Liu Z., Peng Y., Lin Y.: General image-to-image translation with one-shot image guidance. In *ICCV* (2023), pp. 22736–22746. 3

[CRC*20] Chen M., Radford A., Child R., Wu J., Jun H., Luan D., Sutskever I.: Generative pretraining from pixels. In *International conference on machine learning* (2020), PMLR, pp. 1691–1703. 11

[CUYH20] Choi Y., Uh Y., Yoo J., Ha J.-W.: Stargan v2: Diverse image synthesis for multiple domains. In *CVPR* (2020), pp. 8188–8197. 3

[CWC*23] Chen Y., Wang X., Chen X., Zhang Q., Li X., Guo Y., Wang J., Wang F.: Uv volumes for real-time rendering of editable free-view human performance. In *CVPR* (2023), pp. 16621–16631. 2, 3, 8, 10

[CXM*20] Chu M., Xie Y., Mayer J., Leal-Taixé L., Thuerey N.: Learning temporal coherence via self-supervision for gan-based video generation. *TOG 39*, 4 (2020), 75–1. 8

[CZS17] Chu C., Zhmoginov A., Sandler M.: Cyclegan, a master of steganography. *arXiv preprint arXiv:1712.02950* (2017). 3

[DLS*19] Dong H., Liang X., Shen X., Wu B., Chen B.-C., Yin J.: Fw-gan: Flow-navigated warping gan for video virtual try-on. In *ICCV* (2019), pp. 1161–1170. 2, 3

[DPS15] De Paoli C., Singh K.: Secondskin: sketch-based construction of layered 3d models. *TOG 34*, 4 (2015), 1–10. 3

[ESO18] Esser P., Sutter E., Ommer B.: A variational u-net for conditional appearance and shape generation. In *CVPR* (2018), pp. 8857–8866. 3

[FKMW*23] Fridovich-Keil S., Meanti G., Warburg F. R., Recht B., Kanazawa A.: K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR* (2023), pp. 12479–12488. 3

[FLJ*22] Fu J., Li S., Jiang Y., Lin K.-Y., Qian C., Loy C. C., Wu W., Liu Z.: Stylegan-human: A data-centric odyssey of human generation. In *ECCV* (2022), Springer, pp. 1–19. 3

[GBH23] Grigorev A., Black M. J., Hilliges O.: Hood: Hierarchical graphs for generalized modelling of clothing dynamics. In *CVPR* (2023), pp. 16965–16974. 3

[GKJ*21] Garbin S. J., Kowalski M., Johnson M., Shotton J., Valentin J.: Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 14346–14355. 11

[GPAM*14] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.: Generative adversarial nets. *NIPS 27* (2014). 3

[GRH*12] Guan P., Reiss L., Hirshberg D. A., Weiss A., Black M. J.: Drape: Dressing any person. *TOG 31*, 4 (2012), 1–10. 2

[GWL*23] Gao Q., Wang Y., Liu L., Liu L., Theobalt C., Chen B.: Neural novel actor: Learning a generalized animatable neural representation for human actors. *TVCG* (2023). 3

[HJA20] Ho J., Jain A., Abbeel P.: Denoising diffusion probabilistic models. *NIPS 33* (2020), 6840–6851. 3

[HLX*21] Habermann M., Liu L., Xu W., Zollhoefer M., Pons-Moll G., Theobalt C.: Real-time deep dynamic characters. *TOG 40*, 4 (2021), 1–16. 2, 4

[IZZE17] Isola P., Zhu J.-Y., Zhou T., Efros A. A.: Image-to-image translation with conditional adversarial networks. In *CVPR* (2017), pp. 1125–1134. 3

[KGE*21] Kappel M., Golyanik V., Elgharib M., Henningson J.-O., Seidel H.-P., Castillo S., Theobalt C., Magnor M.: High-fidelity neural human motion transfer from monocular video. In *CVPR* (2021), pp. 1541–1550. 3

[KKLD23] Kerbl B., Kopanas G., Leimkühler T., Drettakis G.: 3d gaussian splatting for real-time radiance field rendering. *TOG 42*, 4 (2023). 3

[KLB*23] Kuang Z., Luan F., Bi S., Shu Z., Wetzstein G., Sunkavalli K.: Palettenerf: Palette-based appearance editing of neural radiance fields. In *CVPR* (2023), pp. 20691–20700. 5, 6

[KLF*24] Kwon Y., Liu L., Fuchs H., Habermann M., Theobalt C.: Deliffas: Deformable light fields for fast avatar synthesis. *NIPS 36* (2024). 3

[KRKG24] Karthikeyan A., Ren R., Kant Y., Gilitschenski I.: Avatarone: Monocular 3d human animation. In *WACV* (2024), pp. 3647–3657. 3

[KZW*15] Kwok T.-H., Zhang Y.-Q., Wang C. C., Liu Y.-J., Tang K.: Styling evolution for tight-fitting garments. *IEEE transactions on visualization and computer graphics 22*, 5 (2015), 1580–1591. 3

[KZZ*23] Kang M., Zhu J.-Y., Zhang R., Park J., Shechtman E., Paris S., Park T.: Scaling up gans for text-to-image synthesis. In *CVPR* (2023), pp. 10124–10134. 3

[LCY*23] Liu J.-W., Cao Y.-P., Yang T., Xu Z., Keppo J., Shan Y., Qie X., Shou M. Z.: Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. In *ICCV* (2023), pp. 18483–18494. 3

[LHR*21] Liu L., Habermann M., Rudnev V., Sarkar K., Gu J., Theobalt C.: Neural actor: Neural free-view synthesis of human actors with pose control. *TOG 40*, 6 (2021), 1–16. 2, 3

[LKLR23] Luiten J., Kopanas G., Leibe B., Ramanan D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713* (2023). 3

[LLK19] Liang J., Lin M., Koltun V.: Differentiable cloth simulation for inverse problems. *NIPS 32* (2019). 2

[LMR*15] Loper M., Mahmood N., Romero J., Pons-Moll G., Black M. J.: SMPL: A skinned multi-person linear model. *SIGGRAPH Asia 34*, 6 (Oct. 2015), 248:1–248:16. 6

[LPM*19] Liu W., Piao Z., Min J., Luo W., Ma L., Gao S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV* (2019), pp. 5904–5913. 3

[LTT*20] Li C., Tang M., Tong R., Cai M., Zhao J., Manocha D.: P-cloth: interactive complex cloth simulation on multi-gpu systems using dynamic matrix assembly and pipelined implicit integrators. *TOG 39*, 6 (2020), 1–15. 2

[LZWL23] Li Z., Zheng Z., Wang L., Liu Y.: Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. *arXiv preprint arXiv:2311.16096* (2023). 2

[LZZ*22] Lin S., Zhang H., Zheng Z., Shao R., Liu Y.: Learning implicit templates for point-based clothed human modeling. In *ECCV* (2022), Springer, pp. 210–228. 2

[MBRS*21] Martin-Brualla R., Radwan N., Sajjadi M. S., Barron J. T., Dosovitskiy A., Duckworth D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 7210–7219. 11

[MST*20] Mildenhall B., Srinivasan P. P., Tancik M., Barron J. T., Ramamoorthi R., Ng R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV* (2020). 5

[MST*21] Mildenhall B., Srinivasan P. P., Tancik M., Barron J. T., Ramamoorthi R., Ng R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM 65*, 1 (2021), 99–106. 2, 3

[MWJ12] Meng Y., Wang C. C., Jin X.: Flexible shape control for automatic resizing of apparel products. *Computer-aided design 44*, 1 (2012), 68–76. 3

[MYR*20] Ma Q., Yang J., Ranjan A., Pujades S., Pons-Moll G., Tang S., Black M. J.: Learning to dress 3d people in generative clothing. In *CVPR* (2020), pp. 6469–6478. 2

[NMK*06] Nealen A., Müller M., Keiser R., Boxerman E., Carlson M.: Physically based deformable models in computer graphics. In *Computer graphics forum* (2006), vol. 25, Wiley Online Library, pp. 809–836. 2

[NSLH21] Noguchi A., Sun X., Lin S., Harada T.: Neural articulated radiance field. In *ICCV* (2021), pp. 5762–5772. 3

[NSO12] Narain R., Samii A., O'brien J. F.: Adaptive anisotropic remeshing for cloth simulation. *TOG 31*, 6 (2012), 1–10. 2, 3

[PCPMMN21] Pumarola A., Corona E., Pons-Moll G., Moreno-Noguer F.: D-nerf: Neural radiance fields for dynamic scenes. In *CVPR* (2021), pp. 10318–10327. 3

[PDF*22] Pietroni N., Dumery C., Falque R., Liu M., Vidal-Calleja T. A., Sorkine-Hornung O.: Computational pattern making from 3d garment models. *TOG 41*, 4 (2022), 157–1. 3

[PFSGB20] Pfaff T., Fortunato M., Sanchez-Gonzalez A., Battaglia P. W.: Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409* (2020). 11

[PJBM22] Poole B., Jain A., Barron J. T., Mildenhall B.: Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022). 11

[PLPM20] Patel C., Liao Z., Pons-Moll G.: Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *CVPR* (2020), pp. 7365–7375. 2

[PMJ*22] Pan X., Mai J., Jiang X., Tang D., Li J., Shao T., Zhou K., Jin X., Manocha D.: Predicting loose-fitting garment deformations using bone-driven motion networks. In *SIGGRAPH* (2022), pp. 1–10. 3

[PZX*21] Peng S., Zhang Y., Xu Y., Wang Q., Shuai Q., Bao H., Zhou X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR* (2021), pp. 9054–9063. 2, 3, 7

[RDN*22] Ramesh A., Dhariwal P., Nichol A., Chu C., Chen M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125 1*, 2 (2022), 3. 11

[RPG*21] Ramesh A., Pavlov M., Goh G., Gray S., Voss C., Radford A., Chen M., Sutskever I.: Zero-shot text-to-image generation. In *International conference on machine learning* (2021), Pmlr, pp. 8821–8831. 11

[SCS*22] Saharia C., Chan W., Saxena S., Li L., Whang J., Denton E. L., Ghasemipour K., Gontijo Lopes R., Karagol Ayan B., Salimans T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems 35* (2022), 36479–36494. 11

[SLT*19a] Siarohin A., Lathuilière S., Tulyakov S., Ricci E., Sebe N.: Animating arbitrary objects via deep motion transfer. In *CVPR* (2019), pp. 2377–2386. 3

[SLT*19b] Siarohin A., Lathuilière S., Tulyakov S., Ricci E., Sebe N.: First order motion model for image animation. *NIPS 32* (2019). 3

[SOC19] Santesteban I., Otaduy M. A., Casas D.: Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 355–366. 2

[STOC21] Santesteban I., Thuerey N., Otaduy M. A., Casas D.: Self-supervised collision handling via generative 3d garment models for virtual try-on. In *CVPR* (2021), pp. 11763–11773. 2

[TTG*21] Tretschk E., Tewari A., Golyanik V., Zollhöfer M., Lassner C., Theobalt C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV* (2021), pp. 12959–12970. 3

[TTM*22] Tewari A., Thies J., Mildenhall B., Srinivasan P., Tretschk E., Yifan W., Lassner C., Sitzmann V., Martin-Brualla R., Lombardi S., Simon T., Theobalt C., Niessner M., Barron J. T., Wetzstein G., Zollhöfer M., Golyanik V.: Advances in neural rendering. *Computer Graphics Forum 41*, 2 (2022), 703–735. doi:https://doi.org/10.1111/cgf.14507. 2

[TWL*18] Tang M., Wang T., Liu Z., Tong R., Manocha D.: I-cloth: Incremental collision handling for gpu-based interactive cloth simulation. *TOG 37*, 6 (2018), 1–10. 2

[WCPM18] Wang T. Y., Ceylan D., Popovic J., Mitra N. J.: Learning a shared shape space for multimodal garment design. *arXiv preprint arXiv:1806.11335* (2018). 3

[WCS*22] Weng C.-Y., Curless B., Srinivasan P. P., Barron J. T., Kemelmacher-Shlizerman I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR* (2022), pp. 16210–16220. 3

[WLL*21] Wang P., Liu L., Liu Y., Theobalt C., Komura T., Wang W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689* (2021). 3

[WLZ*18] Wang T.-C., Liu M.-Y., Zhu J.-Y., Liu G., Tao A., Kautz J., Catanzaro B.: Video-to-video synthesis. *arXiv preprint arXiv:1808.06601* (2018). 3

[WSH19] Wolff K., Sorkine-Hornung O.: Wallpaper pattern alignment along garment seams. *TOG 38*, 4 (2019), 1–12. 3

[WWYW20] Wu L., Wu B., Yang Y., Wang H.: A safe and fast repulsion method for gpu-based cloth self collisions. *TOG 40*, 1 (2020), 1–18. 2

[XBS*22] Xiang D., Bagautdinov T., Stuyck T., Prada F., Romero J., Xu W., Saito S., Guo J., Smith B., Shiratori T., et al.: Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *TOG 41*, 6 (2022), 1–15. 2, 4

[XFM22] Xu T., Fujita Y., Matsumoto E.: Surface-aligned neural radiance fields for controllable 3d human synthesis. In *CVPR* (2022), pp. 15883–15892. 2, 3, 8, 10

[XPC*23] Xiang D., Prada F., Cao Z., Guo K., Wu C., Hodgins J., Bagautdinov T.: Drivable avatar clothing: Faithful full-body telepresence with dynamic clothing driven by sparse rgb-d input. In *SIGGRAPH Asia 2023 Conference Papers* (2023), pp. 1–11. 2, 4

[YBZ*22] Yang B., Bao C., Zeng J., Bao H., Zhang Y., Cui Z., Zhang G.: Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *European Conference on Computer Vision* (2022), Springer, pp. 597–614. 11

[YLT*21] Yu A., Li R., Tancik M., Li H., Ng R., Kanazawa A.: Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5752–5761. 11

[YZZ*19] Yu T., Zheng Z., Zhong Y., Zhao J., Dai Q., Pons-Moll G., Liu Y.: Simulcap: Single-view human performance capture with cloth simulation. In *CVPR* (2019), pp. 5504–5514. 2

[ZCF*13] Zhou B., Chen X., Fu Q., Guo K., Tan P.: Garment modeling from a single image. In *Computer graphics forum* (2013), vol. 32, Wiley Online Library, pp. 85–91. 3

[ZSZS19] Zablotskaia P., Siarohin A., Zhao B., Sigal L.: Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139* (2019). 3

[ZWCM21] Zhang M., Wang T. Y., Ceylan D., Mitra N. J.: Dynamic neural garments. *TOG 40*, 6 (dec 2021). 2, 3, 4, 5, 6, 7, 8, 10

[ZWF*19] Zhou Y., Wang Z., Fang C., Bui T., Berg T.: Dance dance generation: Motion transfer for internet videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (2019), pp. 0–0. 2