DressRecon: Freeform 4D Human Reconstruction from Monocular Video

Jeff Tan, Donglai Xiang, Shubham Tulsiani, Deva Ramanan, Gengshan Yang*
Carnegie Mellon University, USA

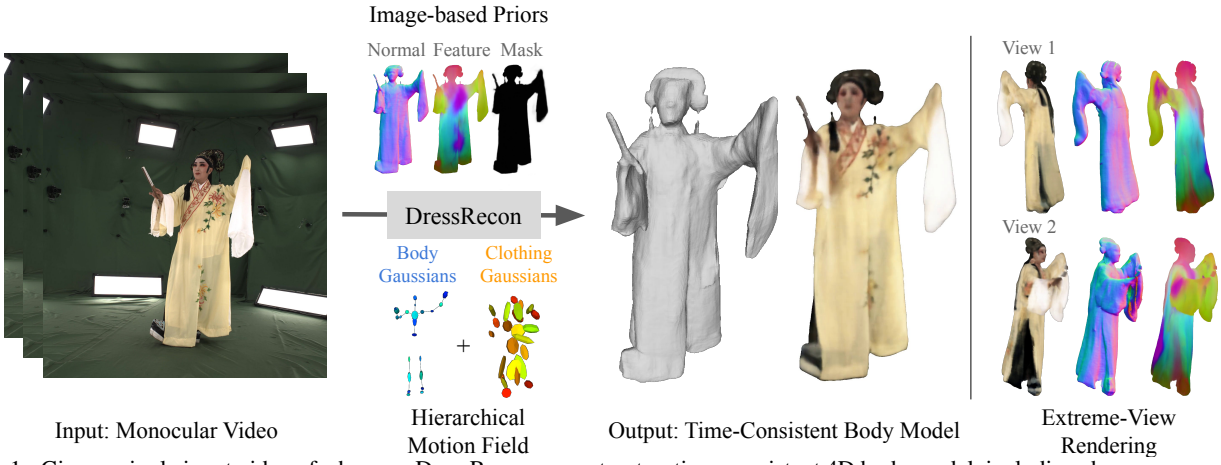

Figure 1. Given a single input video of a human, DressRecon reconstructs a time-consistent 4D body model, including shape, appearance, time-varying body articulations, as well as extremely loose clothing deformation or accessory objects. We propose a hierarchical bag-of-bones deformation model that allows body and clothing motion to be separated. We leverage image-based priors such as human body pose, surface normals, and optical flow to make optimization more tractable. The resulting neural fields can be extracted into time-consistent meshes, or further optimized as explicit 3D Gaussians for high-fidelity interactive rendering.

Abstract

We present a method to reconstruct time-consistent human body models from monocular videos, focusing on extremely loose clothing or handheld object interactions. Prior work in human reconstruction is either limited to tight clothing with no object interactions, or requires calibrated multi-view captures or personalized template scans which are costly to collect at scale. Our key insight for high-quality yet flexible reconstruction is the careful combination of generic human priors about articulated body shape (learned from large-scale training data) with video-specific articulated "bag-of-bones" deformation (fit to a single video via test-time optimization). We accomplish this by learning a neural implicit model that disentangles body versus clothing deformations as separate motion model layers. To capture subtle geometry of clothing, we leverage image-based priors such as human body pose, surface normals, and optical flow during optimization. The resulting neural fields can be extracted into time-consistent meshes, or further optimized as explicit 3D Gaussians for high-fidelity interactive rendering. On datasets with highly challenging clothing deformations and object interactions, DressRecon yields higher-fidelity 3D reconstructions than prior art. Project page: https://jefftan969.github.io/dressrecon/

1. Introduction

We aim to reconstruct animatable dynamic human avatars from videos of people wearing loose clothing or interacting with objects, such as in-the-wild *monocular* videos recorded on a phone or from the Internet. High-quality reconstructions in this setting traditionally require calibrated multi-view captures [40, 63], which are costly to obtain.

From only a single viewpoint, recovering freely-deforming humans with arbitrary topology is highly under-constrained, and thus prior works often rely on domain-

*Corresponding author: jefftan@andrew.cmu.edu

specific constraints which struggle to support loose clothing. Template-based human reconstruction [18, 19, 62] requires personalized scanned templates, which works well for a single instance but cannot reconstruct unseen clothing and body shapes and clothing. Methods that regress 3D surfaces from a single image [60, 61] can produce high-quality geometry at observed regions, but the results are inconsistent across frames and sometimes fail to produce coherent body shapes. Human-specific methods [16, 24, 54] can achieve high quality on tight clothing, but often use a fixed human skeleton or parametric body template and thus cannot handle extreme deformations outside the body. More broadly, generic methods for humans and animals [65, 66] can support arbitrary deformations, but often produce lower quality results than human-specific methods.

This paper presents DressRecon, which reconstructs freeform 4D humans with loose clothing and handheld objects from monocular videos. Our key insight is the careful combination of generic human-level priors about articulated body shape (learned from large-scale training data) with video-specific articulated "bag-of-bones" clothing models (fit to a single video via test-time optimization). We accomplish this by learning a neural implicit model that disentangles body and clothing deformations as separate motion layers. To capture subtle geometry of clothing, we leverage image-based priors such as masks, normals, and body pose during optimization. When the goal is shape reconstruction, we extract time-consistent meshes from the optimized neural fields. Otherwise, to enable high-quality interactive rendering, we propose a refinement stage that converts our implicit neural body into 3D Gaussians while maintaining the motion field design. On datasets with highly challenging clothing and object deformations, DressRecon yields higher-fidelity 3D reconstructions than prior art.

2. Related Work

Humans from multi-view or depth. With sufficient information as input, multi-view methods [9, 13, 26, 37, 40, 45, 58] can reconstruct human shape and appearance of very high fidelity, but the reliance on a dense capture studio limits their applicability at a consumer level. Depth-based methods [10, 59, 72] follow the seminal DynamicFusion work [43] to integrate human shape from a monocular depth stream into a canonical space with the help of a deformation model. However, their application scenarios are also limited because they require specialized depth sensors.

Monocular human reconstruction. Monocular RGB-based reconstruction is challenging due to the 3D ambiguity of a monocular input. Early work [2, 15, 28, 57] aims to reconstruct 3D human keypoints or skeletal poses using a deformable human model [27, 39]. Compared with sparse keypoints, reconstructing dense human surfaces is even more challenging, especially when clothing is considered. Trained on ground truth 3D scans, pixel-aligned implicit functions [34, 47, 61] regress clothed human surfaces from a monocular image, but their output on a video tends to be less temporally coherent. Another line of work aims to reconstruct dynamic human shapes from video input, using a deformable human model [16, 22, 54] or pre-scanned personalized templates [19, 25, 62] and often achieving significant speedups [20, 23, 33]. Generic human models (e.g. SMPL) help resolve monocular 3D ambiguity, but without a personalized clothed template, few works can handle dynamic clothing that does not closely follow body motion. Our method introduces a novel representation that not only leverages human-specific model priors, but also simultaneously enjoys the flexibility to handle loose garments. ReLoo [17] is a concurrent work that applies a two layer deformation model to account for loose garments.

Table 1. **Related work** in monocular 3D body reconstruction. (1)Methods based on human body and pose models. (2)General methods for humans and animals. Dense: Dense deformation fields. Bob: Bag-of-bones. H: Human body and pose priors. F: Optical flow. N: Surface normal. ϕ: Features. Our method combines the best of human-specific and general methods by fitting a flexible motion model initialized from off-the-shelf 3D human poses, using dense image-based priors.

	Method	Motion model	Prior	Input
(1)	ECON [61]	N.A.	H,N	Image
	NeuMan [24]	Skeleton	H	Video
	Vid2Avatar [16]	Skeleton	H	Video
	SelfRecon [22]	Skeleton+Dense	H,N	Video
	HumanNeRF [54]	Skeleton+Dense	H	Video
(2)	MagicPony [56]	Skeleton	ϕ	Image
	LASR [65]	Bob	F	Video
	BANMo [67]	Bob	F,ϕ	Video
	RAC [68]	Skeleton+Dense	F,ϕ	Video
	DressRecon (Ours)	Hierarchical Bob	H,F,N,ϕ	Video

Monocular nonrigid 3D reconstruction. Non-rigid structure from motion (NRSfM) methods [4] reconstruct non-rigid 3D shapes from 2D point trajectories in a class-agnostic way. However, due to the oversimplified motion model and the difficulties in estimating long-range correspondences [49], they do not work well for videos with challenging deformations. Recent work applies differentiable rendering to reconstruct articulated objects from videos [46, 55, 65, 66] or images [14, 29, 56, 71]. However, they cannot reconstruct challenging body articulations and large deformations beyond the body, due to the lack of a flexible motion representation and sufficient measurement signals. As shown in Tab. 1, we introduce a hierarchical bag-of-bones motion model that is capable of representing the deformation of loose garments and accessories, fitted using rich signals from pretrained vision models such as human body pose, surface normals, and optical flow.

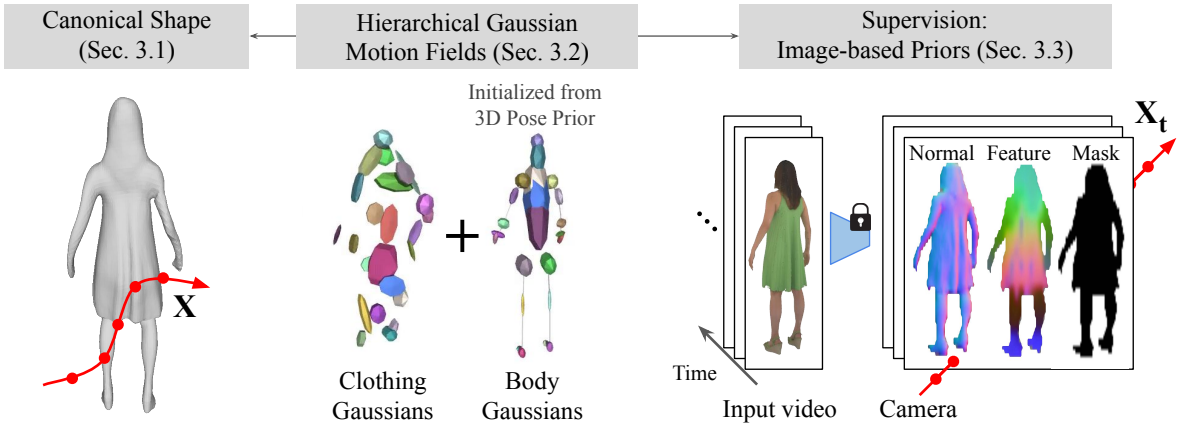

Figure 2. **Method Overview**: We represent 3D humans in loose clothing as temporally consistent 4D neural fields (Sec. 3.1). Central to our approach is a flexible motion representation that captures fine-grained clothing deformations as well as limb motions, while effectively utilizing domain-specific priors such as 3D human body pose (Sec. 3.2). We perform video-specific optimization that fits this model to dense image-based priors via differentiable rendering (Sec. 3.3). After optimization, our neural implicit surface can be extracted into a time-consistent mesh via marching cubes, or converted into explicit 3D Gaussians for high-fidelity interactive rendering (Sec. 3.4).

3. Method

Our goal is to reconstruct time-varying 3D humans in loose clothing from in-the-wild monocular videos (Fig. 2). We represent humans with clothing as 4D neural fields and perform per-video optimization with differentiable rendering (Sec. 3.1). Key to our approach is a hierarchical motion model (Sec. 3.2) capable of representing large limb motions as well as clothing and object deformations. We leverage image-based priors (Sec. 3.3) such as body pose, surface normals, and optical flow to make optimization more stable and tractable. The resulting neural fields can be extracted into time-consistent meshes via marching cubes, or converted into explicit 3D Gaussians for high-fidelity interactive rendering (Sec. 3.4).

3.1. Preliminary: Consistent 4D Neural Fields

To represent a time-varying 3D human, we construct a time-invariant canonical shape that is warped by a time-varying deformation field.

Canonical shape. We represent the body shape as a neural signed distance field in the canonical space, with the following properties: signed distance d, color $\mathbf{c}$, and universal features $\boldsymbol{\phi}$. The canonical fields are defined as

$$(d, \boldsymbol{\phi}) = \mathbf{MLP}_{\text{SDF}}(\mathbf{X}), \tag{1}$$

$$\mathbf{c}_t = \mathbf{MLP}_{\text{color}}(\mathbf{X}, \boldsymbol{\omega}_t), \tag{2}$$

where $\mathbf{X}$ is a 3D point in canonical space and $\boldsymbol{\omega}_t$ is a time-varying appearance code specific to each frame.

Space-time warpings. We represent time-varying motion using continuous 3D deformation fields. A forward deformation field $\mathcal{W}(t)^+ : \mathbf{X} \rightarrow \mathbf{X}_t$ maps a canonical 3D point to time t. During volume rendering, rays at time t are traced back to the canonical space using a backward deformation field $\mathcal{W}(t)^- : \mathbf{X}_t \rightarrow \mathbf{X}$. We use a 3D cycle loss $\mathcal{L}_{\text{cyc}}$ to ensure that $\mathcal{W}(t)^+ \circ \mathcal{W}(t)^-$ is close to identity [35, 67].

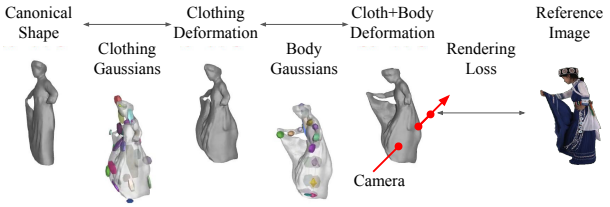

Figure 3. **Visualization of two-layer deformation**. The body and clothing deformation layers each contribute separate types of motion. In this sequence, the clothing Gaussians deform the woman's dress to be larger, while the body Gaussians move her right arm forward. During forward warping, we start from the canonical shape (left), and first apply the forward warp described by clothing Gaussians, then the forward warp described by body Gaussians. The same process happens in reverse during backward warping.

Volume rendering. Neural fields can be optimized via differentiable volume rendering [41], which renders images and minimizes reconstruction errors (e.g. photometric loss). To provide additional supervision on geometry and motion, we augment the training data with additional signals obtained from off-the-shelf networks, detailed in Sec. 3.3.

3.2. Hierarchical Gaussian Motion Fields

In monocular 4D reconstruction, it is challenging to find a motion representation that is both sufficiently flexible and easy to optimize. Recent methods are either not flexible enough to model dynamic structures outside the body [22], or struggle to robustly reconstruct dynamic motions at high quality [65]. We introduce hierarchical motion fields to strike a balance between flexibility and robustness.

Bag-of-bones skinning deformation. Our motion model is inspired by deformation graphs and its extension to Gaussian blend skinning models [3, 50, 65]. The idea is to use the motion of B bones (defined as 3D Gaussians, typically $B = 25$) to drive the canonical geometry's motion. Each Gaussian maintains a time-varying trajectory of its 3D centers $\boldsymbol{\mu}_t \in \mathbb{R}^{T\times 3}$ and orientations $\mathbf{V}_t \in \mathbb{R}^{T\times 3}$ over T frames, as well as axis-aligned scales $\boldsymbol{\Lambda} \in \mathbb{R}^3$ that are time-invariant. Given the 3D Gaussians, a dense forward deformation field can be computed by blending the $\mathbf{SE}(3)$ transformations of Gaussians with forward skinning weights $\mathbf{W}^+$. Similarly, a dense backward deformation field is produced by blending with backward skinning weights $\mathbf{W}_t^-$:

$$\mathbf{X}_t = \mathcal{W}^+(\mathbf{X}, t) = \left(\sum_{b=1}^{B} \mathbf{W}^{+,b}\mathbf{G}_t^b \left(\mathbf{G}^b\right)^{-1}\right)\mathbf{X} \quad (3)$$

$$\mathbf{X} = \mathcal{W}^-(\mathbf{X}_t, t) = \left(\sum_{b=1}^{B} \mathbf{W}_t^{-,b}\mathbf{G}^b \left(\mathbf{G}_t^b\right)^{-1}\right)\mathbf{X}_t \quad (4)$$

Here $\mathbf{G}$ and $\mathbf{G}_t$ are the $\mathbf{SE}(3)$ transformations of the canonical and time t Gaussians, respectively. Forward skinning weights $\mathbf{W}^+ \in \mathbb{R}^b$ are computed using the Mahalanobis distance from $\mathbf{X}$ to each canonical Gaussian $\mathbf{G}$. We use a coordinate MLP to refine the weights (similar to [67]), and use a negative softmax such that farther Gaussians are assigned a lower weight. In the same way, backward skinning weights $\mathbf{W}_t^- \in \mathbb{R}^{T\times b}$ are computed using the Mahalanobis distance from $\mathbf{X}_t$ to each time t Gaussian $\mathbf{G}_t$, followed by MLP refinement.

This bag-of-bones representation can represent large non-rigid deformations due to its flexibility, but can be challenging to optimize. For example, most Gaussians can get concentrated in a local region, which limits the ability to deform the other parts of the target. Carefully initializing the Gaussians and spatially distributing them during optimization can help avoid such bad local minima. Our key idea is to divide the Gaussians into body and clothing layers, which can be initialized and regularized separately.

Body Gaussians are intended to represent skeletal motions of the target. With recent advances in human and animal body pose [15, 42], 3D joint locations can be robustly estimated from images and used to initialize the body Gaussian trajectories. This allows body Gaussians to start from a close-to-optimal solution and get locally refined throughout differentiable rendering. The resulting body Gaussians exhibit less temporal jitter than the single-frame predictor, and are better aligned to physical bone locations.

Clothing Gaussians are intended to represent free-form deformations not explained by body Gaussians, such as cloth deformation and the motion of handheld objects. To encourage that clothing Gaussians only deform structures outside the scope of body Gaussians, we add a regularization term to minimize the impact of clothing Gaussians:

$$\mathcal{L}_{\text{cl}} = \left\|\mathcal{W}^+_{\text{cloth}}(\mathbf{X}, t) - \mathbf{X}\right\|^2 \quad (5)$$

Compositional two-layer deformation. The final deformation fields are the composition of body and clothing layer deformations (Fig. 3), each with about 25 Gaussian bones. During forward warping we apply the clothing deformation before the body deformation, and during backward warping we perform the reverse:

$$\mathcal{W}^+(t) = \mathcal{W}^+_{\text{body}}(t) \circ \mathcal{W}^+_{\text{cloth}}(t) \quad (6)$$

$$\mathcal{W}^-(t) = \mathcal{W}^-_{\text{cloth}}(t) \circ \mathcal{W}^-_{\text{body}}(t) \quad (7)$$

We optimize the body and clothing Gaussians jointly. To encourage body and clothing Gaussians to be well-distributed in 3D space, we use a Sinkhorn divergence loss $\mathcal{L}_{\text{sink}}$ [12] to match the spatial distribution of Gaussians with the body shape. The Sinkhorn divergence is computed between 1k random points on the canonical rest surface, and 3D points on the Gaussians of each deformation layer.

With proper initialization and regularization, body and clothing motion can be properly disentangled. In Fig. 3, the clothing Gaussians deform the dress while the body Gaussians deform the woman's arm. On the supplementary webpage, we show video examples where body and clothing motion is properly decomposed by two-layer deformation.

3.3. Optimization with Image-Based Priors

Optimizing time-varying 3D geometry from monocular videos is challenging due to its under-constrained nature. Recent advances in surface normals [11], optical flow [51, 64], image features [5, 44], and zero-shot segmentation [32] provide additional interpretations of raw pixel values. This knowledge is not only generic, but also highly correlated with the geometry and motion of the underlying scene, making it suitable for our reconstruction task. We introduce an optimization routine that uses foundational image-based priors as supervision to make the problem tractable.

Surface normals. Without multi-view inputs, it is challenging to distinguish shape from appearance. For example, detailed structures such as clothing wrinkles can just as easily be painted as colors on a flat surface, leading to inaccurate surface geometry. To counteract this, we use normal estimators [31] trained on large datasets to provide a signal to improve the geometry. We can take spatial derivatives of signed distance d with respect to $\mathbf{X}_t$ to compute the surface normal of a point $\mathbf{X}_t$ in deformed space. We normalize the rendered and estimated surface normals and compute a normal loss as $\mathcal{L}_2$ error between them. Similar to prior work on neural surface reconstruction [70], we also compute an

eikonal loss $\mathcal{L}_{\text{eik}}$ to regularize the neural surface.

$$\mathbf{n} = \text{normalize}(\nabla d(\mathcal{W}^{-}(\mathbf{X}_t, t))) \tag{8}$$

$$\mathcal{L}_{\mathbf{n}} = \|\mathbf{n} - \mathbf{n}^*\|^2 = 2 - 2\langle \mathbf{n}, \mathbf{n}^* \rangle \tag{9}$$

$$\mathcal{L}_{\mathbf{eik}} = \big\|\text{norm}(\nabla d(\mathcal{W}^{-}(\mathbf{X}_t, t))) - 1\big\| \tag{10}$$

Normals with numerical gradients. Most prior work uses analytical gradients (e.g. auto-diff) to compute normals of signed distance fields. However, these are computed within an infinitesimally small neighborhood of $\mathbf{X}_t$ and suffer from noise in both the estimated backward warping fields and signed distances [7]. This leads to unstable optimization when dealing with deformable objects. To avoid this, we compute normals by numerical gradients [36] with a fixed 1mm step size during optimization.

Normals with eikonal filtering. Although numerical normal computation works well on static scenes, it is more challenging in deformable scenes where the warping field's influence can cause $\|\mathbf{X}_t + \delta - \mathbf{X}_t\|$ to be very different from $\|\mathcal{W}^{-}(\mathbf{X}_t + \delta) - \mathcal{W}^{-}(\mathbf{X}_t)\|$. For example, the hand and waist might be close in deformed space but far in canonical space, causing exploding gradients due to a large change in signed distance gradient over a small neighborhood. To avoid this problem, we clip the normal direction to 0 after Eq. 8 whenever the gradient magnitude exceeds some threshold, in our case $\|\nabla d\| > 10$.

Optical flow. We use optical flow [51, 64] to learn the non-rigid deformation and relative camera transform between two frames. We compute 3D scene flow vectors by backward warping deformed points to canonical space, then forward-warping to another timestamp. We use the camera matrix to project 3D flow vectors into 2D, and compute $\mathcal{L}_2$ error between rendered flow $\mathbf{f}$ and estimated flow $\mathbf{f}^*$, $L_{\mathbf{f}} = \|\mathbf{f} - \mathbf{f}^*\|$. Here, $|t' - t| = \{1, 2, 4, 8\}$:

$$\mathbf{f}_{\text{3D}}(\mathbf{X}_t, t \to t') = \mathcal{W}^{+}(\mathcal{W}^{-}(\mathbf{X}_t, t), t') - \mathbf{X}_t \tag{11}$$

Universal features. Deep neural features are useful for registering pixels to a 3D model [38, 66], while allowing better convergence at textureless regions or under deformation. Prior work relies on category-specific image features, but we find DINOv2 [44]) to be a robust and universal feature descriptor that works well for clothing and accessories. We choose the small DINOv2 model with registers, as it produces fewer peaky feature artifacts [8]. We obtain pixel-level features from DINOv2's patch descriptors by evaluating DINOv2 on an image pyramid, averaging features across pyramid levels, and reducing the dimension to 16 via PCA [1]. We compute feature loss as $\mathcal{L}_2$ error between rendered and estimated features, $\mathcal{L}_{\boldsymbol{\phi}} = \|\boldsymbol{\phi} - \boldsymbol{\phi}^*\|^2$.

Zero-shot segmentation. Inspired by shape-from-silhouette [52], we use image segmentation to carve out the 3D boundary of the target. We leverage the foundational 2D segmentation model SAM [32] and its extension to tracking [69] to predict accurate silhouettes of humans with clothing and accessories. We pass different prompts according to different scenarios we aim to reconstruct, such as "human wearing cloth" and "human holding an object". We compute silhouette loss as the $\mathcal{L}_2$ error between rendered and estimated silhouettes, $\mathcal{L}_{\mathbf{s}} = \|\mathbf{s} - \mathbf{s}^*\|^2$.

Losses. Our final loss is a weighted sum of reconstruction and regularization terms. Loss weights λ are searched once and kept across all experiments.

$$\mathcal{L}_{\text{rec}} = \lambda_{\mathbf{c}}\mathcal{L}_{\mathbf{c}} + \lambda_{\mathbf{f}}\mathcal{L}_{\mathbf{f}} + \lambda_{\mathbf{n}}\mathcal{L}_{\mathbf{n}} + \lambda_{\boldsymbol{\phi}}\mathcal{L}_{\boldsymbol{\phi}} + \lambda_{\mathbf{s}}\mathcal{L}_{\mathbf{s}} \tag{12}$$

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{eik}}\mathcal{L}_{\text{eik}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{sink}}\mathcal{L}_{\text{sink}} + \lambda_{\text{cl}}\mathcal{L}_{\text{cl}} \tag{13}$$

3.4. Refinement with 3D Gaussians

Representation. Neural SDFs are ideal for extracting surfaces, but can be difficult to optimize as adding new geometry requires making global changes. In light of this, we introduce a refinement procedure that replaces the canonical shape representation with 3D Gaussians [30] while keeping the two-layer motion model as is. To render an image, we warp Gaussians forward from canonical space to time t (Eq. 17) and call the differentiable Gaussian rasterizer.

Initialization. We use 40k Gaussians, each parameterized by 14 values, including its opacity, RGB color, center location, orientation, and axis-aligned scales. Gaussians are initialized on the surface of the neural SDF with isotropic scaling. To initialize the color of each Gaussian, we query the canonical color MLP (Eq. 2) at its center.

Optimization. We update both the canonical 3D Gaussian parameters and the motion fields by minimizing

$$\mathcal{L}_{\text{rec}} = \lambda_{\mathbf{c}}\mathcal{L}_{\mathbf{c}} + \lambda_{\mathbf{f}}\mathcal{L}_{\mathbf{f}} + \lambda_{\mathbf{s}}\mathcal{L}_{\mathbf{s}} \tag{14}$$

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{sink}}\mathcal{L}_{\text{sink}}. \tag{15}$$

Notably, the 3D cycle loss $\mathcal{L}_{\text{cyc}}$ can be dropped since rasterization does not require computing backward warps.

4. Experiments

We evaluate DressRecon's ability to reconstruct both 3D shape and appearance given challenging monocular videos. Video results are available on the supplementary webpage.

4.1. Datasets

Dynamic clothing and accessories. To evaluate DressRecon's ability to reconstruct dynamic clothing and objects, we select 14 sequences from DNA-Rendering [6] with challenging cloth deformation and/or handheld objects (e.g. playing a cello, swinging a cloth, waving a brush). As DNA-Rendering does not provide ground-truth meshes, we compute pseudo-ground-truth 3D meshes by using all 48 available cameras to optimize a separate NeuS2 [53] instance at each timestep. To overcome the limited viewpoint range of each individual camera, we assemble turntable

monocular videos by rendering these per-frame NeuS2 instances along a smooth 360-degree camera trajectory.

Avatars from casual videos. We also evaluate DressRecon's ability to recover high-fidelity human avatars from casual turntable videos. We evaluate our method on ActorsHQ [21] and select subsets of the first 4 sequences for evaluation, each about 200 frames. As ActorsHQ cameras have small fields of view and often do not cover the whole body, we colorize the provided ground-truth meshes and render turntable monocular videos with 360° of camera rotation.

4.2. Results

Reconstructing dynamic clothing and accessories. Tab. 2 reports the 3D chamfer distance (cm, ↓) for reconstructing dynamic clothing and handheld objects, evaluated across 14 DNA-Rendering sequences. We compare with Vid2Avatar [16], BANMo [67], RAC [68], and ECON [61], and show qualitative results in Fig. 4. The project page contains corresponding video results. DressRecon reconstructs finer details and more accurate body shape than prior art, and is able to handle challenging scenarios such as the tip of the cello (image 1), the hair tassels (image 2), and the detailed cloth wrinkles on the martial arts uniform (image 4).

Reconstructing avatars from casual videos. Tab. 4 reports 3D chamfer distance (cm, ↓) and F-score at $\{1, 2, 5\}$-cm thresholds for recovering avatars from turntable videos, evaluated across 4 ActorsHQ sequences. We compare with Vid2Avatar [16] and show qualitative results in Fig. 5. DressRecon performs on par with Vid2Avatar in tight clothing scenarios, and reconstructs higher-fidelity geometry on sequences with challenging clothing such as dresses.

Rendering dynamic clothing and accessories. Tab. 3 reports the RGB PSNR (↑), SSIM (↑), LPIPS (↓), and mask IoU (↑) on test views by holding out every 8-th training view. We compare against Vid2Avatar [16], BANMo [67], and RAC [68], and show qualitative results in Fig. 7. The project page contains extensive video results. DressRecon produces more accurate renderings than prior art.

4.3. Diagnostics

3D Gaussian refinement. In Tab. 6, we show results from optimizing a neural implicit model from scratch, a 3D Gaussian model from scratch, and a 3D Gaussian model initialized from a neural implicit model. The same computational budget is allocated to all three experiments. The highest rendering quality is achieved with neural implicit optimization followed by 3D Gaussian refinement. This suggests that the neural implicit model helps produce a good initialization of shape and deformation, making it easier for 3DGS to converge to better local optima.

Choice of deformation model. In Tab. 5, we swap our hierarchical two-layer deformation model with several alternatives in the literature. Swapping to a skeleton+dense warping field [22, 54], skeleton alone [68], or bag-of-bones alone [67] reduces the geometry quality. Alternative defor-

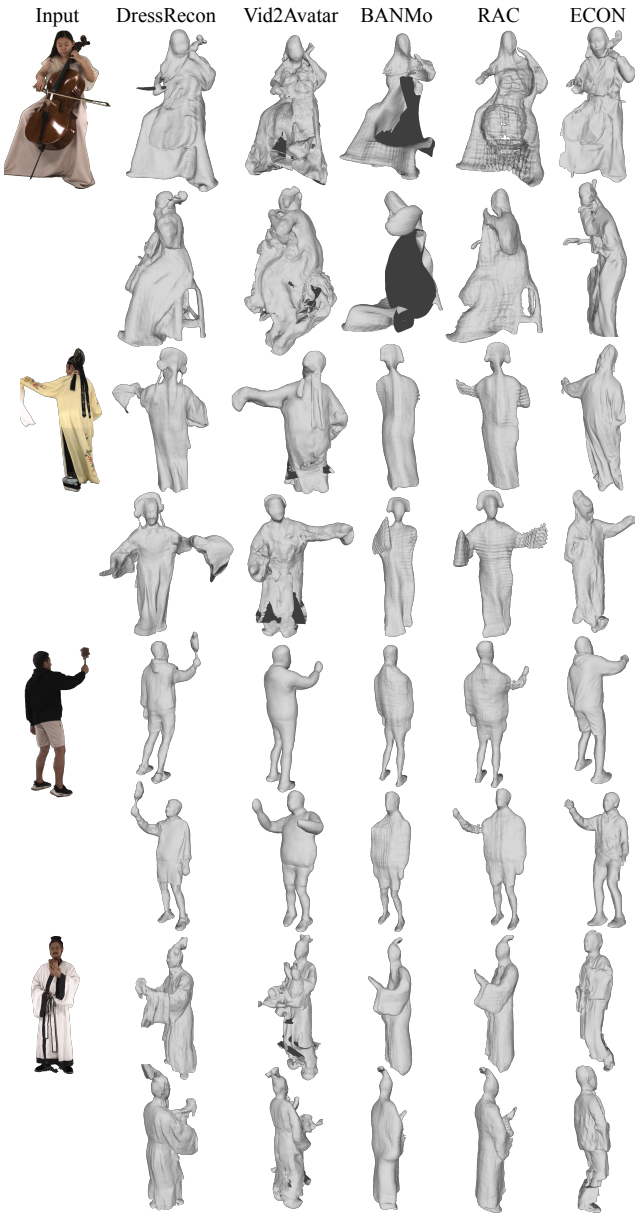

Figure 4. **3D reconstruction results on DNA-Rendering.** We demonstrate DressRecon's ability to reconstruct challenging sequences with large cloth deformation. DressRecon's predictions align well with the image evidence, even in the presence of rapid clothing and object deformations. Vid2Avatar often outputs spurious shape artifacts and is unable to reconstruct challenging structures, such as the white cloth (row 2), brown brush (row 3), and detailed sleeves (row 4). BANMo and RAC produce hollow cellos on the first row, and tend to output over-smoothed surfaces for the other cases. ECON produces highly detailed textures, but it performs the worst numerically (Tab. 2) as the outputs often have an incorrect overall shape (e.g. Row 1). We encourage readers to view the video results on the supplementary webpage.

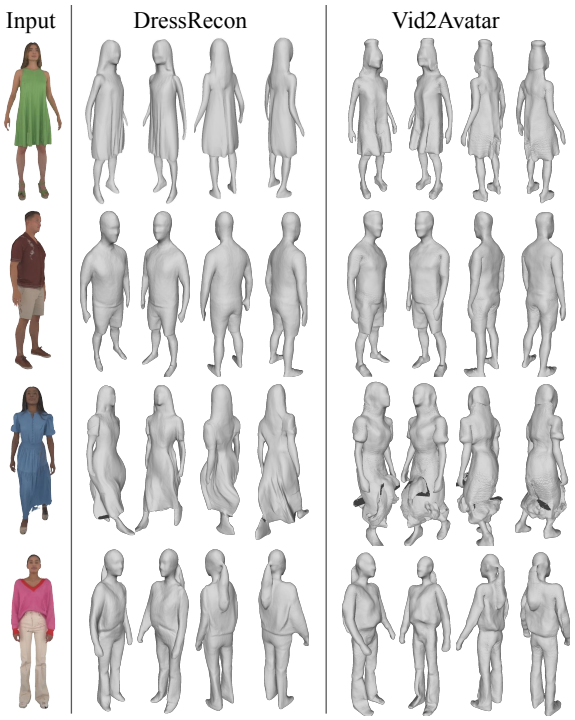

Figure 5. **3D reconstruction results on ActorsHQ.** DressRecon is on par with Vid2Avatar for standard clothing (Rows 2 and 4), and higher fidelity than Vid2Avatar for loose clothing (Rows 1 and 3). Vid2Avatar's reconstructed skirts often contain shape artifacts. We attribute DressRecon's improved performance to its flexible shape and deformation representation, which is capable of representing non-standard geometry and deformation.

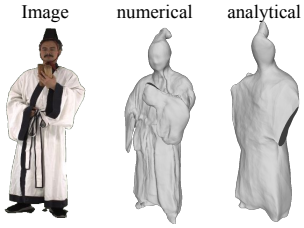

Figure 6. **Qualitative ablation of numerical normals**. We show the difference between optimizing with numerical and analytical normals. Using analytical normals causes training to be unstable, resulting in a flat shape with no surface detail. The quality of surface details is reduced when normal loss is disabled (Tab. 5).

mation models are also less interpretable, as skeleton-only and bag-of-bones do not separate body and clothing motion.

Choice of image-based priors. In Tab. 5, we run the optimization routine and remove one of the image-based priors each time. Without mask loss, the surface geometry has an incorrect overall structure. Without normal loss, the reconstructed surface has lower detail. Without flow loss, the shape is less sensible and camera optimization is less stable.

Choice of normal supervision. In Fig. 6, we show the benefit of using normal loss with numerical gradients. With analytical gradients, shape optimization becomes unstable.

Table 2. **3D reconstruction metrics on DNA-Rendering sequences.** We evaluate 3D chamfer distance (cm, ↓) on fourteen DNA-Rendering sequences with challenging clothing deformation or handheld objects. DressRecon outperforms all baselines, and is the **best** or second-best method on all sequences.

Sequence	DressRecon (Ours)	Vid2Avatar [16]	BANMo [67]	RAC [68]	ECON [61]
0008_01	7.300	**6.786**	7.768	8.112	9.420
0047_01	8.542	11.216	**7.808**	9.106	17.102
0047_12	**6.064**	9.334	6.914	6.618	7.541
0102_02	5.421	7.812	**5.131**	7.278	10.181
0113_06	**6.872**	8.517	8.362	8.391	11.282
0121_02	**5.520**	6.478	7.453	6.926	9.334
0123_02	**6.725**	8.343	7.418	9.683	12.108
0128_04	**7.803**	8.184	8.913	10.005	11.569
0133_07	**6.194**	6.314	7.017	7.449	7.320
0152_01	6.437	7.465	6.815	**6.170**	8.735
0166_04	**4.356**	4.608	5.969	6.286	6.562
0188_02	**5.403**	5.887	6.341	5.829	9.026
0206_04	**7.555**	8.392	8.404	9.644	9.987
0239_01	5.559	**5.503**	6.503	7.831	8.026
Average	**6.411**	7.489	7.201	7.809	9.871

Table 3. **Rendering metrics on DNA-Rendering sequences.** We evaluate RGB PSNR (↑), SSIM (↑), LPIPS (↓), and mask IoU (↑), averaged across fourteen DNA-Rendering sequences with challenging clothing deformation or handheld objects. DressRecon outperforms all baselines, particularly when 3D Gaussian refinement is used to improve the rendering quality.

Method	PSNR	SSIM	LPIPS	Mask IoU
DressRecon	22.03	0.9375	0.1059	0.9544
w/ 3DGS refinement	**22.27**	**0.9506**	**0.0860**	**0.9786**
Vid2Avatar [16]	19.61	0.8948	0.1167	0.7931
BANMo [67]	19.78	0.9341	0.1257	0.8988
RAC [68]	19.89	0.9351	0.1157	0.9044

Table 4. **3D reconstruction metrics on ActorsHQ sequences.** We evaluate 3D chamfer distance (cm, ↓) and F-score at $\{1, 2, 5\}$-cm thresholds (%, ↑) on the four ActorsHQ sequences shown in Fig. 5. DressRecon outperforms Vid2Avatar on most sequences.

	DressRecon				Vid2Avatar			
Sequence	CD	F@5	F@2	F@1	CD	F@5	F@2	F@1
a1s1	3.212	94.74	**72.46**	**48.99**	**3.204**	**98.22**	69.69	39.25
a2s1	**1.838**	**99.96**	**92.72**	**62.14**	2.891	97.59	79.35	40.65
a3s1	**2.647**	**97.15**	**79.91**	**51.38**	4.376	90.93	56.83	30.33
a4s1	2.247	**98.88**	85.02	57.00	**2.039**	98.82	**89.26**	**63.84**
Average	**2.486**	**97.68**	**82.53**	**54.88**	3.128	96.39	73.78	43.52

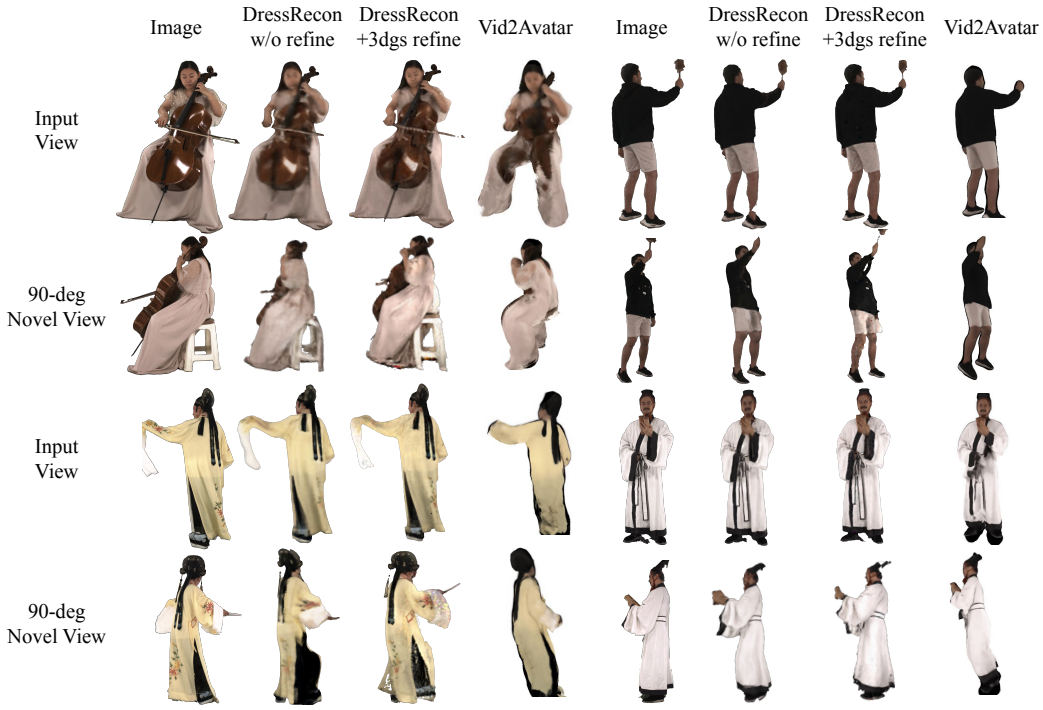

Figure 7. **RGB rendering results on DNA-Rendering.** For each sequence, we show DressRecon's and Vid2Avatar's renderings at both the input view and a 90-degree novel view. DressRecon's renderings are shown with and without 3D Gaussian refinement. We find (similar to Tab. 3) that refinement significantly improves the textures, especially the flowers on the yellow dancer's sleeve. Vid2Avatar's renderings are less detailed, and fail to accurately depict structures that substantially deviate from the body, such as the cello and white stool.

Table 5. **Ablation study for 3D reconstruction.** We ablate the importance of motion field representation and choice of image-based priors, by evaluating 3D chamfer distance (cm, ↓) and F-score at $\{1, 2, 5\}$-cm thresholds (%, ↑) on 14 DNA-Rendering sequences. DressRecon performs worse after switching motion representations (skeleton-only [16], bag-of-bones [67], skeleton+dense [22]) and after removing any image-based prior.

Sequence	CD	F@5	F@2	F@1
DressRecon	**6.411**	**81.16**	**47.66**	**25.62**
skeleton-only	7.340	78.10	44.55	22.67
bag-of-bones	6.942	80.19	47.21	25.49
skeleton+dense	7.526	76.33	41.58	21.91
w/o mask	10.647	65.73	33.61	17.42
w/o normal	7.206	77.70	43.46	23.53
w/o flow	7.094	79.03	45.72	24.41
w/o pose	6.938	78.87	46.21	25.11
w/o feat	6.829	79.25	46.75	25.34

5. Discussion

We present DressRecon, which reconstructs humans with loose clothing and accessory objects from monocular videos. DressRecon uses hierarchical bag-of-bones deformation to model clothing and body deformation separately, and leverages off-the-shelf priors such as masks and surface normals to make optimization more tractable. To improve the rendering quality, we introduce a refinement stage that converts the implicit neural body into 3D Gaussians.

Table 6. **Ablation study for Gaussian refinement.** We ablate the impact of 3D Gaussian refinement, by evaluating RGB PSNR (↑), SSIM (↑), LPIPS (↓), and mask IoU (↑) on 14 DNA-Rendering sequences. We perform experiments where only an implicit SDF is optimized, where 3D Gaussians are optimized without initializing from an SDF, and where a neural SDF is used to initialize 3D Gaussians. The best rendering quality is obtained by initializing 3D Gaussians from an SDF.

Sequence	PSNR	SSIM	LPIPS	Mask IoU
Implicit-only	22.03	0.9375	0.1059	0.9544
3DGS-only	21.31	0.9455	0.0939	0.9737
Implicit→3DGS	**22.27**	**0.9506**	**0.0860**	**0.9786**

Limitations. DressRecon requires sufficient view coverage to reconstruct a complete human, and cannot hallucinate unobserved body parts. It also has no understanding of cloth deformation physics. As a result, clothing may deform unnaturally if we reanimate with novel body motion. We leave reanimating human-cloth and human-object interactions as future work. Moreover, specifying inaccurate segmentation, e.g. by passing the wrong prompt to SAM [32], could result in failure to reconstruct some details.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 5

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 2

[3] Aljaž Božič, Pablo Palafox, Michael Zollhöfer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. *CVPR*, 2021. 4

[4] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2000. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 4

[6] Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-centric Rendering. *arXiv*, 2023. 5

[7] Aditya Chetan, Guandao Yang, Zichen Wang, Steve Marschner, and Bharath Hariharan. Accurate differential operators for hybrid neural fields. *arXiv preprint arXiv:2312.05984*, 2023. 5

[8] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 5

[9] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 2

[10] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4): 1–13, 2016. 2

[11] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10786–10796, 2021. 4

[12] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019. 4

[13] 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

[14] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoints without keypoints. In *ECCV*, 2020. 2

[15] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 4

[16] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2Avatar: 3D Avatar Reconstruction from Videos in the Wild via Self-supervised Scene Decomposition. *CVPR*, 2023. 2, 6, 7, 8

[17] Chen Guo, Tianjian Jiang, Manuel Kaufmann, Chengwei Zheng, Julien Valentin, Jie Song, and Otmar Hilliges. Reloo: Reconstructing humans dressed in loose garments from monocular video in the wild. In *European conference on computer vision (ECCV)*, 2024. 2

[18] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. LiveCap: Real-time Human Performance Capture from Monocular Video. *ACM TOG*, 2019. 2

[19] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. DeepCap: Monocular Human Performance Capture Using Weak Supervision. *CVPR*, 2020. 2

[20] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. 2024. 2

[21] 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 TOG*, 2023. 6

[22] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. 2, 3, 6, 8

[23] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. *arXiv*, 2022. 2

[24] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, pages 402–418. Springer, 2022. 2

[25] Yue Jiang, Marc Habermann, Vladislav Golyanik, and Christian Theobalt. Hifecap: Monocular high-fidelity and expressive capture of human performances. *arXiv preprint arXiv:2210.05665*, 2022. 2

[26] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 41(1):190–204, 2017. 2

[27] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 2

[28] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[29] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2

[30] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 5

[31] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models, 2024. 4, 2

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4, 5, 8, 2

[33] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. 2024. 2

[34] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 49–67. Springer, 2020. 2

[35] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 3, 1

[36] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *CVPR*, pages 8456–8465, 2023. 5

[37] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. *arXiv preprint arXiv:2311.16096*, 2023. 2

[38] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In *ICCV*, 2021. 5

[39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 2

[40] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. *3DV*, 2024. 1, 2

[41] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[42] Tanmay Nath*, Alexander Mathis*, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie W Mathis. Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature Protocols*, 2019. 4

[43] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, pages 343–352, 2015. 2

[44] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4, 5, 2

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

[46] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*, 2020. 2

[47] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–2314, 2019. 2

[48] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, 2021. 1

[49] Peter Sand and Seth Teller. Particle video: Long-range motion estimation using point trajectories. In *IJCV*, 2008. 2

[50] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM SIGGRAPH 2007 papers*. 2007. 4

[51] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 4, 5

[52] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *SIGGRAPH 2008*. 2008. 5

[53] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5

[54] 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 *CVPR*, pages 16210–16220, 2022. 2, 6

[55] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *arXiv preprint arXiv:2107.10844*, 2021. 2

[56] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3d animals in the wild. 2023. 2

[57] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 2

[58] 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. 2

[59] 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. *arXiv preprint arXiv:2310.05917*, 2023. 2

[60] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. *CVPR*, 2022. 2

[61] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. *CVPR*, 2023. 2, 6, 7

[62] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. MonoPerfCap: Human Performance Capture from Monocular Video. *ACM TOG*, 2018. 2

[63] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4K4D: Real-Time 4D View Synthesis at 4K Resolution. *arXiv*, 2023. 1

[64] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. In *NeurIPS*, 2019. 4, 5, 2

[65] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. LASR: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021. 2, 3, 4, 1

[66] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. In *NeurIPS*, 2021. 2, 5

[67] Gengshan Yang, Minh Vo, Neverova Natalia, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 2, 3, 4, 6, 7, 8

[68] Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, and Deva Ramanan. Reconstructing Animatable Categories from Videos. *CVPR*, 2023. 2, 6, 7, 1

[69] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 5

[70] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 2021. 4, 1

[71] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *CVPR*, 2021. 2

[72] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 910–919, 2017. 2

DressRecon: Freeform 4D Human Reconstruction from Monocular Video

Supplementary Material

6. Video Results

Please see the attached webpage for video results.

7. Implementation Details

7.1. Consistent 4D Neural Fields

Signed distance fields. We initialize canonical signed distance fields as a sphere with radius 0.1m. Following standard practice, we apply positional encodings to all 3D points ($L_{xyz} = 10$) and timestamps ($L_t = 6$) before passing into MLPs. The appearance code $\boldsymbol{\omega}_t$ has 32 channels.

After $\mathbf{MLP}_{\text{SDF}}$ computes the signed distance d at a 3D point, we convert the signed distance to a volumetric density $\sigma \in [0, 1]$ for volume rendering. Similar to VolSDF [70], this is done using the cumulative Laplace distribution $\sigma = \Gamma_\beta(d)$, where β is a global learnable scalar parameter that controls the solidness of the object, approaching zero for solid objects. This representation allows us to extract a mesh as the zero level-set of the SDF.

Cycle consistency regularization. Given a forward warping field $\mathcal{W}^+(t) : \mathbf{X} \rightarrow \mathbf{X}_t$ and a backward warping field $\mathcal{W}^-(t) : \mathbf{X}_t \rightarrow \mathbf{X}$, we introduce a cycle consistency term, similar to NSFF [35]. A sampled 3D point in camera coordinates should return to its original location after passing through a backward and forward warping:

$$\mathcal{L}_{\text{cyc}} = \sum_{\mathbf{X}_t} \|\mathcal{W}^+(\mathcal{W}^-(\mathbf{X}_t, t), t) - \mathbf{X}_t\|_2^2 \tag{16}$$

7.2. Hierarchical Gaussian Motion Fields

Bag-of-bones skinning deformation. Our motion model uses the motion of B bones (defined as 3D Gaussians, typically $B = 25$) to drive the motion of canonical geometry. Given 3D Gaussians, we compute dense 3D motion fields by blending the $\mathbf{SE}(3)$ transformations of canonical Gaussians with skinning weights $\mathbf{W}$:

$$\mathbf{X}_t = \mathcal{W}^+(\mathbf{X}, t) = \left(\sum_{b=1}^{B} \mathbf{W}^{+,b} \mathbf{G}_t^b \left(\mathbf{G}^b\right)^{-1} \right) \mathbf{X} \tag{17}$$

$$\mathbf{X} = \mathcal{W}^-(\mathbf{X}_t, t) = \left(\sum_{b=1}^{B} \mathbf{W}_t^{-,b} \mathbf{G}^b \left(\mathbf{G}_t^b\right)^{-1} \right) \mathbf{X}_t \tag{18}$$

where $\mathbf{G}_t^+$ are forward warps from canonical to time t Gaussians, $\mathbf{G}_t^-$ are backward warps from time t to canonical Gaussians, and $\mathbf{W}^+$ are forward skinning weights.

Similar to SCANimate [48] and LASR [65], we define a forward skinning weight function $\mathcal{S}^+ : \mathbf{X} \rightarrow \mathbb{R}^B$ which computes the normalized influence of each Gaussian bone on a canonical 3D point. At a coarse level, skinning weights are defined as the Mahalanobis distance from $\mathbf{X}$ to the canonical Gaussians:

$$\mathbf{W}_\sigma^+ = (\mathbf{X} - \boldsymbol{\mu})^\top \mathbf{Q} (\mathbf{X} - \boldsymbol{\mu}), \tag{19}$$

where $\boldsymbol{\mu} \in \mathbf{R}^{B\times 3}$ are canonical bone centers, $\mathbf{Q} = \mathbf{V}^\top \boldsymbol{\Lambda} \mathbf{V}$ are canonical bone precision matrices, $\mathbf{V} \in \mathbf{R}^{B\times \mathbf{SO}(3)}$ are canonical bone orientations, and $\boldsymbol{\Lambda}^{B\times 3\times 3}$ are time-invariant axis-aligned diagonal scale matrices.

In addition to a coarse component, we find it helpful to use delta skinning weights to model fine geometry. Delta skinning weights are computed by a coordinate MLP:

$$\mathbf{W}_\Delta^+ = \mathbf{MLP}_{\Delta,+}(\mathbf{X}, t) \in \mathbb{R}^B \tag{20}$$

The final skinning function is a normalized sum of coarse and fine components:

$$\mathbf{W}^+ = \mathcal{S}^+(\mathbf{X}, t) = \text{softmax}(-\mathbf{W}_\sigma^+ - \mathbf{W}_\Delta^+), \tag{21}$$

where the negative sign ensures that faraway Gaussian bones (which have a larger Mahalanobis distance) are assigned a lower skinning weight after softmax.

Backward skinning weights are computed analogously with the time t Gaussians, which have center $\boldsymbol{\mu}_t$, orientation $\mathbf{V}_t$, and time-invariant scale $\boldsymbol{\Lambda}$. We also need the transformation $\mathbf{G}_t^-$ from each time t Gaussian to the canonical Gaussian, as well as the backward skinning $\mathbf{MLP}_\Delta^-$.

7.3. Optimization

Sampling. Due to the expensive per-ray computation in volume rendering, optimization with batch gradient descent is challenging. As a result, previous methods randomly sample entire images [68] to compute the reconstruction terms, leading to small batch sizes (typically 16 images per batch) and noisy gradients. We implement an efficient data-loading pipeline with memory-mapping that allows per-pixel measurements (e.g., RGB, flow, features) to load directly from disk without accessing the full image. This allows loading pixels from significantly more images in a single batch (e.g. 256 images on a GPU).

Hyperparameters. We use the Adam optimizer with learning rate 0.0005. We use 48k iterations of optimization for all experiments. On a single RTX 4090 GPU, it takes about 8 hours to optimize the neural implicit body model and 15 seconds to render each frame. 3D Gaussian refinement is performed for another 48k iterations of optimization, taking about 8 hours to optimize and 0.1 seconds to render each

Table 7. **Summary of losses and loss weights.** Our final loss is a weighted sum of reconstruction terms (color, optical flow, normal, feature, and segmentation) and regularization terms (eikonal, cycle-consistency, gaussian consistency, camera prior, and joint prior).

Loss	Weight	Description
$\mathcal{L}_{\mathbf{c}}$	$\lambda_{\mathbf{c}} = 0.1$	L2 loss, rendered RGB vs. the input image
$\mathcal{L}_{\mathbf{f}}$	$\lambda_{\mathbf{f}} = 0.5$	L2 loss, rendered 2D flow vs. computed flow from VCNPlus [64]
$\mathcal{L}_{\mathbf{n}}$	$\lambda_{\mathbf{n}} = 0.03$	L2 loss, rendered normals vs. computed normals from Sapiens [31]
$\mathcal{L}_{\phi}$	$\lambda_{\phi} = 0.01$	L2 loss, rendered features vs. computed features from DINOv2 [44]
$\mathcal{L}_{\mathbf{s}}$	$\lambda_{\mathbf{s}} = 0.1$	L2 loss, rendered masks vs. computed masks from SAM [32]
$\mathcal{L}_{\text{eik}}$	$\lambda_{\text{eik}} = 0.01$	Encourage numerical gradients of canonical SDF to have unit norm
$\mathcal{L}_{\text{cyc}}$	$\lambda_{\text{cyc}} = 0.05$	Encourage backward and forward warping fields to be inverses
$\mathcal{L}_{\text{gauss}}$	$\lambda_{\text{gauss}} = 0.2$	Sinkhorn divergence between canonical 3D Gaussians and SDF
$\mathcal{L}_{\text{cloth}}$	$\lambda_{\text{cloth}} = 0.1$	Minimize the magnitude of clothing deformation

frame. Our loss weights are described in Tab. 7. At each iteration, we sample 72 images and take 16 pixel samples per image. For training efficiency, input images are cropped to a tight bounding box around the object and resized to 256x256. To prevent floater artifacts from appearing outside the tight crop, 90% of pixel samples are taken from the tight bounding box and 10% of pixel samples are taken from the full un-cropped image.