ICCV #15622

NGD: Neural Gradient Based Deformation for Monocular Garment Reconstruction

Anonymous ICCV submission

Paper ID 15622

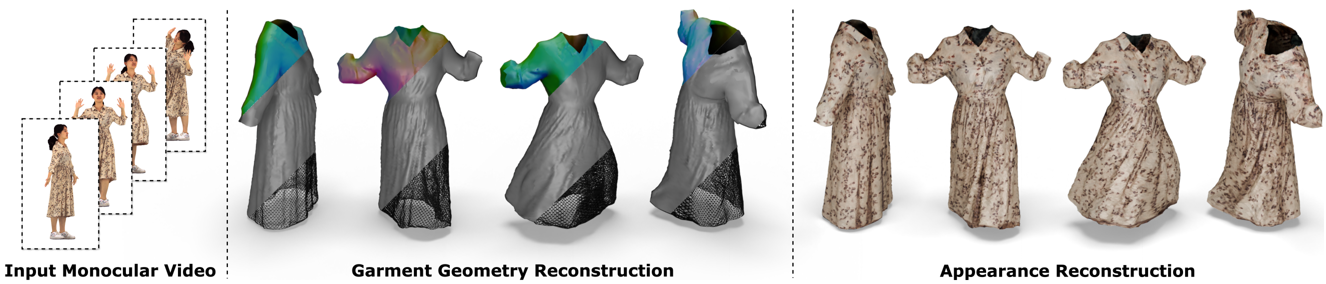

Figure 1. Our method reconstructs high-fidelity garment geometry and appearance from input monocular video.

Abstract

Dynamic garment reconstruction from monocular video is an important yet challenging task due to the complex dynamics and unconstrained nature of the garments. Recent advancements in neural rendering have enabled high-quality geometric reconstruction with image/video supervision. However, implicit representation methods that use volume rendering often provide smooth geometry and fail to model high-frequency details. While template reconstruction methods model explicit geometry, they use vertex displacement for deformation which results in artifacts. Addressing these limitations, we propose NGD, a ***N****eural* ***G****radient-based* ***D****eformation method to reconstruct dynamically evolving textured garments from monocular videos. Additionally, we propose a novel adaptive remeshing strategy for modeling dynamically evolving surfaces like wrinkles and pleats of the skirt, leading to high-quality reconstruction. Finally, we learn dynamic texture maps to capture per-frame lighting and shadow effects. We provide extensive qualitative and quantitative evaluations to demonstrate significant improvements over existing SOTA methods and provide high-quality garment reconstructions.*

1. Introduction

Recent advances in computer vision have enabled large-scale digitization of 3D garments for immersive AR/VR platforms, revolutionizing Social Media, E-Commerce, Gaming, and Entertainment industries. The sheer diversity, complex dynamics, and intricate articulations make garment digitization and modeling significantly challenging. Unlike conventional digital garment creation methods involving artists, which demanded expertise, time, and labor, deep learning has enabled garment digitization from images and videos [28, 33, 34, 37, 47]. Multi-view video inputs are used to obtain high-quality garment digitization, but they often require expensive calibrated multi-camera setups [19, 27, 34, 47, 50], and hence difficult to scale. In comparison, monocular video inputs are easy to acquire and scalable with an abundance of "in the wild" videos available. Nevertheless, garment digitization from monocular video needs to reconstruct the dynamically evolving garment geometry and appearance while addressing the classical challenges like modeling varying garment sizes, non-rigid deformations due to body shapes and poses, and the diverse topology of garments.

Advancements in differentiable rendering have made it possible to achieve high-quality geometry reconstruction from monocular videos. [4, 9, 14, 26, 33, 34]. The existing approaches for garment reconstruction can be divided into implicit surface deformation methods [9, 33] and explicit template deformation methods [4, 26]. SCARF [9] is one of the first works to use implicit surface representation using Neural Radiance Fields (NeRF)[30]; nevertheless, the geometric quality is limited by constraints inherent to volume rendering approaches. REC-MV [33] addresses this limitation by optimizing for both explicit feature curves and implicit garment surfaces. However, the use of implicit representations adds an overhead of surface extraction and the resulting surface is smooth, losing out high-fidelity surface details. Pergamo[4] and Dgarments[26] deform garment templates with SMPL interpolated skinning weights supervised via differentiable rendering. However, these methods rely on a fixed template constraining its ability to model dynamically varying topology, like the pleats of the skirt. Additionally, direct vertex displacement via differentiable rendering causes abrupt, sharp local changes and requires additional regularisers for smoothening undesirable local deformations. This often results in an over-smoothed surface and fails to capture high-frequency details.

To address the above limitations, we develop a *Neural Gradient Based Deformation* method to reconstruct dynamic garments from input monocular video. Our method models appearance and geometry separately, as learning them together might result in appearance being corrected to compensate for geometric inaccuracies and vice versa. We propose a novel deformation parameterization that decomposes surface deformations into a frame-invariant component representing the global shape and a frame-dependent component modeling the pose-specific local surface deformations of the base garment mesh. Specifically, the deformation parameterization adopts NJF [1] to model garment reconstruction, which learns a local Jacobian field defined on the garment surface followed by a Poisson-solve to predict the global garment deformation in canonical space. This addresses the aforementioned limitation of existing template-based methods. These canonical garments are skinned to model garment reconstruction to map to the corresponding input monocular views and optimized via a differentiable renderer [22]. While there are existing approaches that combine NJF with a differentiable renderer, [10], we develop a gradient-based deformation approach to model from the monocular input video. Unlike existing methods that directly optimize with colored images, which can result in inaccuracies due to ambiguities between shadows and textures, we use *diffuse garment images*. Additionally, we design an adaptive remeshing strategy to iteratively increase the mesh resolution in the regions of high-frequency geometrical details. This enables regions with fine details to be modeled by higher mesh resolutions and also freely deform the template to model extremely loose garments. Finally, we learn appearance via dynamic texture maps at each frame to capture lighting and shadow effects. Figure 1 visualize the high-fidelity dynamic textured garment reconstructed by our method from an input monocular video. In summary, our key technical contributions are as follows:

- We propose a novel method to reconstruct dynamically evolving textured garments from monocular videos.
- Our novel deformation parameterization combined with the novel adaptive remeshing enables modeling extremely loose garments with high-frequency details.
- We provide qualitative and quantitative comparisons with existing methods to show significant improvements, especially on loose garments.

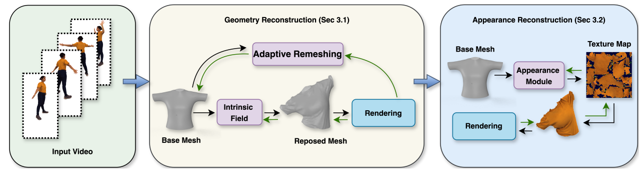

Figure 2. **Method overview:** Given an input video, we reconstruct dynamically evolving textured garment meshes using our Geometry and Appearance Reconstruction module.

2. Related Works

A large number of existing methods attempted clothed human reconstruction from single or multi-view images [15–17, 35, 43, 46, 48] or videos [3, 12, 13, 19, 32, 36, 41, 42], albeit cannot extract garment mesh separately. On the other hand, several existing garment reconstruction methods [5–7, 18, 24, 25, 28, 29, 31, 49] recover garments from monocular image. However, these single-image reconstruction methods require supervised training on a large dataset. Please refer to the supplementary for a detailed discussion of these methods.

Multiview images can recover garments in a self-supervised manner. Diffavatar [27] uses sewing patterns to represent garments and obtain simulation-ready garments from multiview images. Gaussian Garments [34] combines physics simulation with Gaussian splats [20] to obtain physically plausible garments from multiview inputs. The rendering captures fine details down to the level of furs. While multiview reconstruction provides rich garment digitization solutions, the multiview camera setups are generally expensive, hence monocular videos provide a cheap, scalable alternative.

DeepCap [14] is one of the pioneering approaches to reconstructing loose garments from monocular video. However, it considers the first frame as a template, requiring expensive preprocessing including 3D scanning of a clothed human, segmentation, and reconstruction of the garment and human separately. Methods like Pergamo [4] deform garment templates using SMPL-interpolated skinning weights, followed by rendering loss optimization. [9] integrates a parametric body model with [30] representation for garment reconstruction; however, the geometric quality is constrained by NeRF's inherent limitations. REC-MV [33] uses implicit-explicit representation to achieve ge-

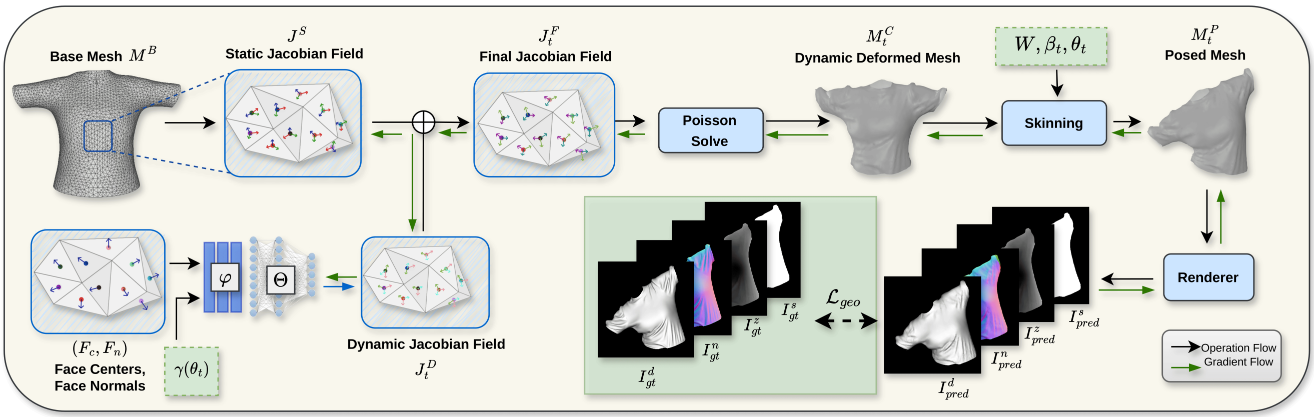

Figure 3. In Geometry Reconstruction Module we introduce a novel deformation parameterization to deform a base mesh M^B to desired target mesh via learning a Jacobian Field guided by differentiable rendering supervision from input monocular video.

ometrically consistent and temporally coherent garment reconstruction. Despite this, they lack detailed textures, and their use of initial implicit garment representations leads to smoothing effects, compromising high-fidelity detail. The recent method, DGarments [26], achieves state-of-the-art performance in geometry reconstruction from monocular video by introducing a multi-hypothesis deformation module. However, they fail to large deformations and struggle with loose clothing.

3. Method

We present NGD, a novel approach for reconstructing dynamically evolving textured garment meshes from given input monocular video. Our method is composed of geometry and appearance reconstruction modules, as shown in Figure 2. As part of our geometry reconstruction module, we introduce a novel deformation parameterization over a base garment mesh to accurately capture and aggregate garment deformations across input frames. This parameterization decomposes deformations into a frame-invariant component representing the global canonical shape and a frame-dependent component modeling the pose-specific local surface deformations of the garment. To further improve geometric fidelity, we also propose a novel Gradient-Based Remeshing Strategy subsubsection 3.1.1, which adaptively refines the mesh resolution in regions exhibiting high curvature thereby facilitating the precise modeling of intricate details, such as wrinkles and folds. Our appearance reconstruction module subsection 3.2 learns garment appearance by learning a frame-invariant base texture map and a frame-dependent dynamic texture map that captures the visual characteristics of the garments.

3.1. Geometric Reconstruction Module

The base garment mesh M^B is a 2-manifold embedded in 3D Euclidean space $\mathbb{R}^3$. Let $\mathbf{V} := \{v_i \in \mathbb{R}^3\}_{i=0}^N$, $\mathbf{F} := \{f_j \in \mathbb{N}^3\}_{j=0}^M$ and $\mathbf{E} := \{e_l \in \mathbb{N}^2\}_{l=0}^L$ be the vertices, faces and edges of the mesh M^B respectively. We separately model the global deformations capturing garment-specific design features (such as collars, and necklines) as well the local dynamic deformations (such as wrinkles) on M^B in T-pose at every time-frame. To achieve this, we find a mapping function Φ_t that transforms the base mesh M^B to a desired mesh $\tilde{M}_t$ in canonical space (T-pose) that captures these dynamic deformations at each time-frame t. This mapping function $\Phi_t : \mathbb{R}^{N\times 3} \rightarrow \mathbb{R}^{N\times 3}$ is approximated by optimizing for Jacobian fields and using Poisson Solve to obtain deformed mesh vertices [1]. However, unlike NJF [1] and TextDeformer [10] which optimizes for a single static mesh, we need to approximate a mapping function Φ_t corresponding to every frame.

Thus, given input video frames $\mathbf{I} = \{I_t\}_{t=0}^T$, the goal is to find the optimal deformation function Φ_t at every frame by solving the following equation in the least square sense:

$$\Phi_t^* = \min_{\Phi_t} \sum_{f_j \in F} |f_j| \, \|\nabla_i \Phi_t - J_j\|^2, \tag{1}$$

where ∇ is the gradient operator. The function Φ_t^* ideally maps the base garment M^B to target mesh $\tilde{M}_t$. The solution to the above equation Equation 1 is obtained by solving a Poisson system [1]. This mapping function Φ_t^* is indirectly estimated by optimizing for the Jacobians J_t^F to obtain the canonical mesh M_t^C, which is the closest approximation of the desired mesh $\tilde{M}_t$.

Intrinsic Deformation Fields: Building on the aforementioned Jacobian Field formulation, we propose a novel deformation parameterization for dynamic garment modeling by splitting J_t^F into two sub-fields, a frame-invariant static

ICCV #15622

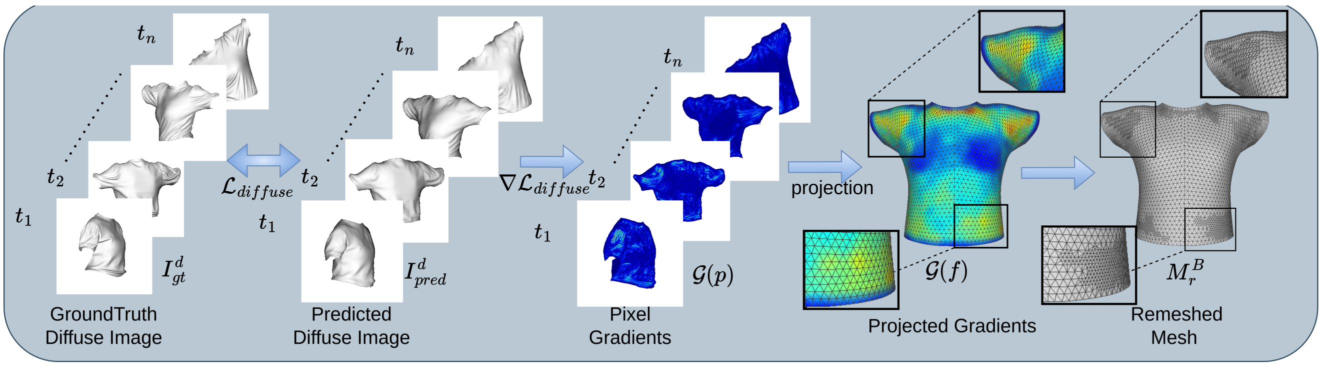

Figure 4. **Overview of our gradient-based adaptive remeshing method:** Performing edge selection, followed by remeshing operations for generating remeshed meshs with high frequency details.

Jacobian Field $J^S \in \mathbb{R}^{M\times3\times3}$, and a frame-specific dynamic Jacobian Field $J_t^D \in \mathbb{R}^{M\times3\times3}$. J^S captures the global garment shape specific to the input video garment style. This static field is defined at each face center of the base mesh, initialized as an identity matrix, and is optimized directly across all frames. The dynamic field J_t^D captures the pose-specific surface deformation at each image frame and is predicted by a neural network f_G.

The Figure 3 shows how these two Jacobian Fields model per-frame deformations in the canonical space. The neural network $f_G = f_\Theta \circ f_\varphi$ is composed of hash-grid encoder f_φ and an MLP f_Θ. At every time-step t, we use face centers F_C, face normals F_N of the reposed static canonical garment mesh, and pose information for conditioning the neural network. Conditioning with the pose defined by the joint angles θ_t prevents overfitting to the input view. We use the PCA (Principle Component Analysis) for encoding the pose parameters as $\gamma(\theta_t)$. More details about the pose encoding are provided in the Suppl. Finally, the MLP takes as input latent encoding of F_N, F_C and $\gamma(\theta_t)$ from f_φ to predict J_t^D. The final Jacobian field is defined as $J_t^F = J^S + J_t^D$. The final Jacobian field J_t^F is solved via the Poisson system to obtain canonical garment M_t^C encompassing both global garments specific as well as local surface deformations.

Skinning Transformation: The canonical garment M_t^C is subsequently skinned to obtain the reposed garment for every time-frame M_t^P defined as follows:

$$M_t^P = S(M_t^C, \beta_t, \theta_t, W) \tag{2}$$

where $S(.)$ is the skinning function, β_t and θ_t are the shape and pose parameters, and W is the garment skinning weights. This reposed mesh is rendered to obtain diffuse and depth images of the garment. The pseudo ground truth extracted from the input images guides the optimization of the Jacobian Field J^S and the neural network f_G parameters via a differentiable renderer. Figure 3 provides a visual overview of our geometry reconstruction module.

Local Minima: The optimization is often trapped in local minima while minimizing the local rendering losses. To address this, we introduce a novel exponentially decaying noise applied to the vertices of the final skinned mesh iteratively. This noise encourages the model to prioritize global geometry in the initial iterations, preventing early overfitting to local details. This heuristic adds no computational overhead while significantly improving the reconstruction quality of loose garments (refer to supplementary for detailed discussion).

Losses: The normal maps I^n_{gt}, part segmentation maps I^s_{gt}, and depth maps I^z_{gt} extracted from input images serve as pseudo-ground truth to optimize the reconstruction module. Instead of using normal maps, we use diffused maps I^d_{gt} obtained by projecting light in the input camera view direction, for supervision. Thus, the rendering loss is defined as:

$$\mathcal{L}_{\text{diffuse}} = \mathcal{H}((I^s_{gt} \odot I^d_{\text{pred}}), I^d_{\text{gt}}) + \mathcal{S}((I^s_{\text{gt}} \odot I^d_{\text{pred}}), I^d_{\text{gt}}) \tag{3}$$

where $\odot$ is the element-wise multiplication, $\mathcal{H}$ is the Huber Loss, $\mathcal{S}$ is the SSIM loss and I^d_{pred} is the diffuse images calculated from the input of the predicted mesh normals.

The regularization loss, $\mathcal{L}_{\text{reg}}$ ensures continuous surface consistency after deformation. The per-triangle Jacobians $J_j \in \mathbb{R}^{3\times3}$ of the final intrinsic field J_t^F is optimized to be close to the identity matrix $I \in \mathbb{R}^{3\times3}$, defined as:

$$\mathcal{L}_{\text{reg}} = \sum_{j=1}^{M} \|J_j - I\|_2^2 \tag{4}$$

Finally, we use a depth supervision loss, $\mathcal{L}_{\text{depth}}$, calculated using the depth-ranking scheme proposed in [38]. A modified segmentation loss $\mathcal{L}_{\text{mask}}$ is used for supervision from segmentation masks (more detail in Suppl.). The total geometric reconstruction loss $\mathcal{L}_{\text{geo}}$ is defined as:

$$\mathcal{L}_{\text{geo}} = \lambda_1\mathcal{L}_{\text{render}} + \lambda_2\mathcal{L}_{\text{mask}} + \lambda_3\mathcal{L}_{\text{reg}} + \lambda_4\mathcal{L}_{\text{depth}} \tag{5}$$

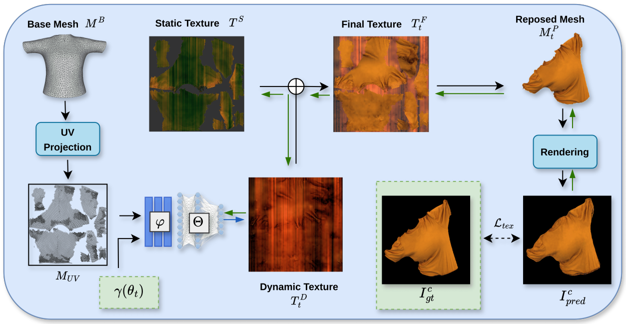

Figure 5. Overview of our appearance reconstruction module.

3.1.1. Gradient Based Adaptive Remeshing

We select a set of edges E_s, based on their gradients from rendering loss $\mathcal{L}_{\text{diffuse}}$ and then apply remeshing operations, as illustrated in Figure 4.

Edge Selection: Out of all edges $\mathbf{E}$ in the base mesh M^B, we select a subset $E_s \subset \mathbf{E}$ for remeshing. The image-space gradient at each pixel p is defined as $\mathcal{G}(p) = \nabla_{I^d_{\text{pred}}(p)} \mathcal{L}_{\text{diffuse}}$ Equation 3 where $I^d_{\text{pred}}(p)$ is the predicted image. These pixel gradients are aggregated over rasterized faces $\Pi(f_j)$ for each face $f_j \in \mathbf{F}$, resulting in per rasterized face gradient values $\mathcal{G}(\Pi_{\text{raster}}(f_j))$. These values are then aggregated over all iterations and projected onto the base mesh M^B, yielding a per-face gradient value: $\mathcal{G}(f_j) = \frac{\Sigma \mathcal{G}(\Pi_{\text{raster}}(f_j))}{|\Pi_{\text{raster}}(f_j)|}$. Next, we select the top quantile of faces $\mathcal{F}_\omega = \{f_j \mid \|\mathcal{G}(f_j)\| \geq \text{quantile}_\omega(\|\mathcal{G}(f_j)\|)\}$ for a percentile ω of triangle face. Subsequently, we prune all faces $\mathcal{F}_\delta = \{f_j \mid L(e_l) \geq \delta_{\text{length}}, \forall e_l \in \mathcal{E}(f_j)\}$ whose edge lengths fall below a certain threshold δ_{length}. The selection threshold and pruning criteria evolve over epochs over a linearly decaying function, ensuring a balance between preserving details and preventing excessive refinement. Finally, we select all edges E_s part of all the final selected faces $\mathcal{F}_\delta$.

Remeshing: Subsequently, we perform edge splitting and edge flipping operations on E_s, adopting the remeshing strategy proposed in [8]. During the remeshing process, it is crucial to handle face flips and degenerate triangles. Finally, we clean up the mesh to remove degenerate faces and merge close vertices. This yields the modified topology base mesh M_r^B. Next, we need to recomputation of all mesh attributes. After remeshing, the mesh attributes are recomputed via k-NN interpolation. The static Jacobian field J^S, Adam optimizer moments m_1, m_2, and skinning weights W are interpolated to ensure smooth training. Please refer to Suppl. for more information.

3.2. Appearance Reconstruction Module

Our goal is to learn a dynamic texture map corresponding to the reposed mesh at each frame. The detailed overview of texture recovery is provided in Figure 5. We obtain the base UV coordinates M_{UV} from M_r^B using [23]. These UV coordinates map the color information from a texture map to the mesh faces. Similar to geometry reconstruction, we learn two texture components. A frame-invariant static texture map $T^S \in \mathbb{R}^{q \times q \times 3}$, and a per-frame dynamic texture map $T_t^D \in \mathbb{R}^{q \times q \times 3}$, where q is the texture image dimension. The static texture T^S is optimized directly, while the dynamic texture T_t^D is predicted by a neural network. At every time-step t, the MLP f_T is conditioned on hash encoded UV coordinates $f_\varphi(M_{UV})$, and pose parameters $\gamma(\theta_t)$ to predict T_t^D. The final Texture map is obtained as $T_t^F = T^S + T_t^D$. For better generalization, we employ a smooth annealing training strategy, inspired by [44], wherein we introduce linearly decaying Gaussian noise to the pose parameters, $\gamma(\theta_t)$. This approach effectively mitigates overfitting and improves generalizability in novel views synthesis. At each iteration, the posed garment mesh M_t^P from the geometry module is rendered with color from texture T_t^F, to produce colored images I^c_{pred}. The static texture T^S and the neural network parameters of f_T are optimized via differentiable rendering with the following two losses: $\mathcal{L}_{\text{col}} = ||(I^s_{gt} \odot I^c_{pred}), I^c_{gt}||_1$ and $\mathcal{L}_{\text{ssim}} = \text{SSIM}(I^s_{gt} \odot I^c_{pred}), I^c_{gt}$.
The final loss is defined as:

$$L_{tex} = \alpha_1 L_{col} + \alpha_2 L_{ssim} \tag{6}$$

4. Experiments & Results

4.1. Implementation Details

Our proposed method is implemented in PyTorch with NVDiffrast [22] as the core differentiable rasterizer. The primary training for our method was conducted on a single NVIDIA RTX 4090 GPU, for both geometry and appearance reconstruction. Each sequence of 100 frames takes approximately 2.5 hours to train including texture recovery. Both modules incorporate a fixed-epoch warm-up phase, during which only the static deformation field J^S and static texture map T^S are optimized. After the warm-up phase, the dynamic deformation field J_t^D and dynamic texture map T_t^D are introduced for joint optimization. Adaptive remeshing is performed at fixed intervals throughout the optimization process.

4.2. Experimental Setup

We evaluate and compare our method against recent State-Of-The-Art (SOTA) approaches on two tasks: 3D surface reconstruction and novel view synthesis. Our evaluation spans five sequences from a modified 4D-Dress [39] dataset, along with two additional datasets [2, 33], selecting two sequences from each to demonstrate robustness. We provide quantitative comparisons for both tasks on the

ICCV
#15622

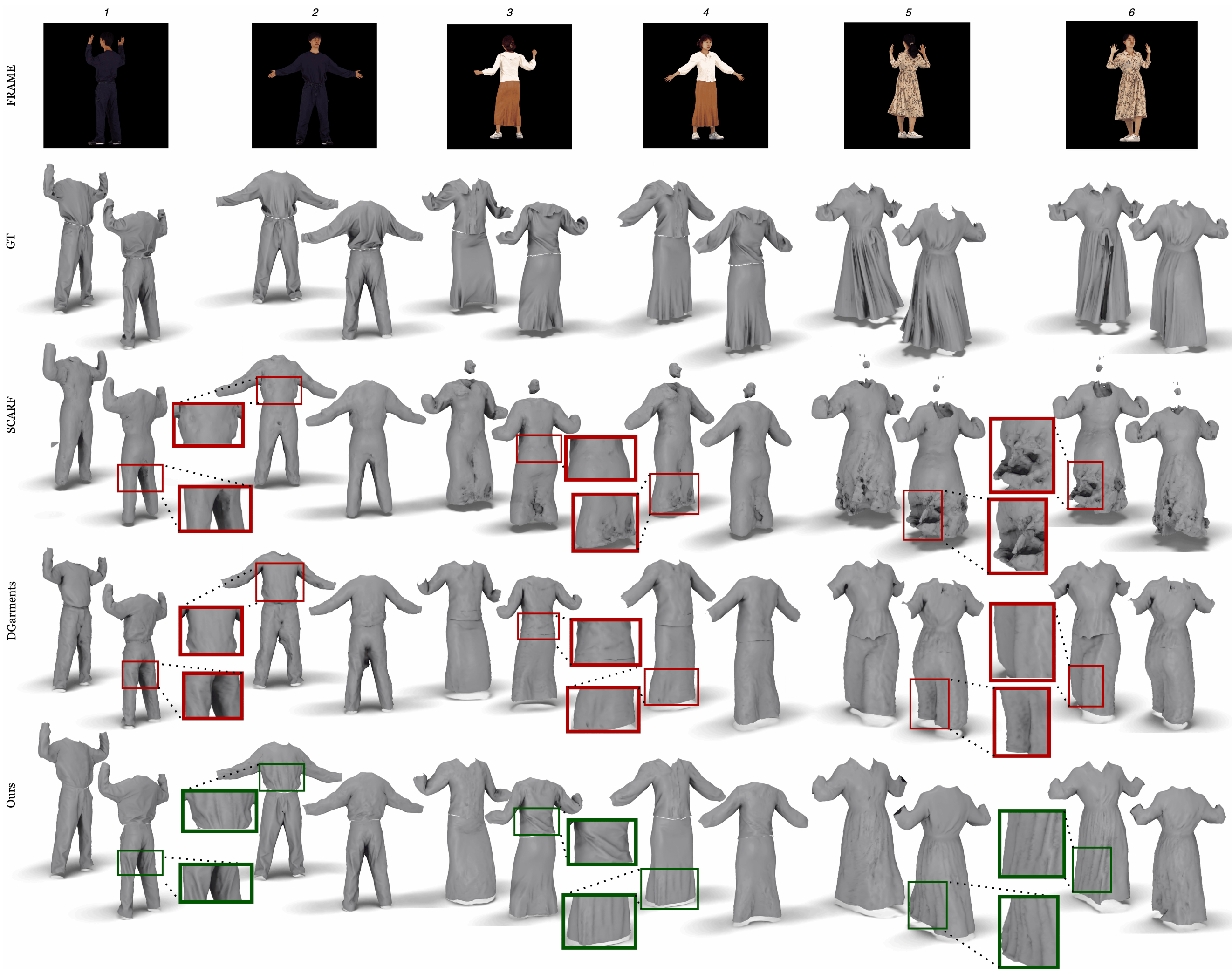

Figure 6. Qualitative comparison where our method faithfully reconstructs high-frequency details like tiny wrinkles and folds, closer to GroundTruth in comparison to SCARF [9] and DGarments [26] on 4D-Dress dataset [39].

Table 1. Quantitative evaluation on geometry reconstruction on 4D-Dress dataset [39] using Chamfer Distance (CD) and Normal Consistency (NC) and comparison with different methods.

	Chamfer Distance $\mathcal{L}_2 \times 10^3 \downarrow$						Normal Consistency $\uparrow$					
Method	123	148	169	185	187	Avg	123	148	169	185	187	Avg
SCARF	8.622	-	6.507	2.423	3.261	5.203	0.915	-	0.872	0.837	0.753	0.844
DGarment	0.076	0.863	0.154	0.431	1.722	0.649	0.904	0.755	0.872	0.856	0.777	0.833
Ours	**0.050**	**0.660**	**0.127**	0.393	**0.923**	**0.431**	**0.934**	**0.766**	**0.891**	**0.879**	**0.794**	**0.853**
w/o remeshing	0.053	0.672	0.129	**0.372**	0.981	0.441	0.932	0.762	0.887	0.878	0.790	0.850
w normals	0.195	0.931	0.278	0.535	1.205	0.554	0.908	0.755	0.866	0.853	0.778	0.832

4D-Dress dataset [39]. Additionally, we provide qualitative comparisons for 4D-Dress dataset for both tasks across all datasets. To assess the effectiveness of our model, we perform comparisons with the following SOTA methods - REC-MV [33], SCARF [9], and DGarment [26]. Finally, we provide extensive ablation studies to analyze our design choices. Please refer Suppl. for Dataset specifications and implementation details.

Figure 7. Qualitative comparison of geometric reconstruction obtained by our method with SCARF [9] and REC-MV [33] on People Snapshot [2] dataset. Our method faithfully reconstructs high-frequency details like tiny wrinkles and folds.

Table 2. Quantitative evaluation on novel view synthesis with PSNR (PR), SSIM (SM), and LPIPS (LS) on different sequences.

Sequence	123			169			185			187		
Method	**PR** ↑	**SM** ↑	**LS** ↓	**PR** ↑	**SM** ↑	**LS** ↓	**PR** ↑	**SM** ↑	**LS** ↓	**PR** ↑	**SM** ↑	**LS** ↓
SCARF	43.02	0.992	0.018	45.01	0.992	0.026	33.82	0.986	0.025	25.32	0.918	0.0828
Ours	**46.78**	**0.998**	**0.008**	**47.91**	**0.996**	**0.014**	**35.21**	**0.990**	**0.017**	**25.85**	**0.948**	**0.0395**

Figure 8. Qualitative comparison of novel view synthesis.

Data Preprocessing: We utilize existing pre-trained vision models to obtain reliable priors. The SMPL pose and shape parameters and the camera estimations are obtained from 4DHumans [11]. Per-frame normal map, depth map, and part-segmentation are recovered using a pre-trained human foundation model Sapiens [21]. Finally, the base garment mesh is obtained using BCNet [18].

4.3. Results

Geometry Reconstruction : Quantitative evaluation, presented in Table 2 (rows [1-3]), demonstrates that our method significantly outperforms the SOTA methods [9, 26] both in terms of Normal Consistency (NC) as well as Chamfer Distance (CD), averaged across all frames of a sequence. We achieve a significantly improved alignment of the reconstructed garment mesh with the ground truth mesh while achieving consistent geometrical characteristics across different frames, leading to substantially lower average CD values & higher average NC values across the sequence as well as average overall sequences across the dataset.

A similar trend is evident in the qualitative evaluation presented in Figure 6. The qualitative differences are more significant for col [3 − 6] which contains loose clothing such as gown, where we outperform the existing methods while effectively mitigating major artifacts as shown in col 5. The qualitative results for additional datasets [2, 33]

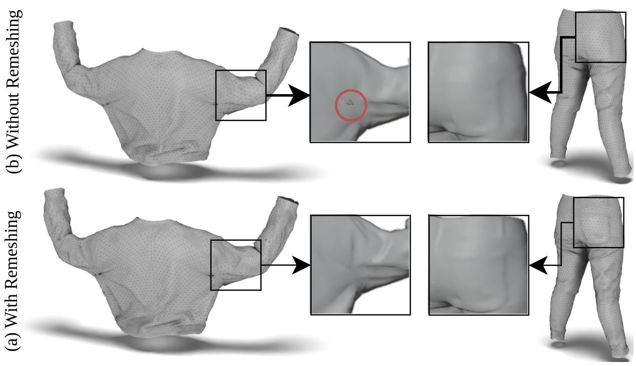

Figure 9. Ablative results on Gradient Based Adaptive Remeshing.

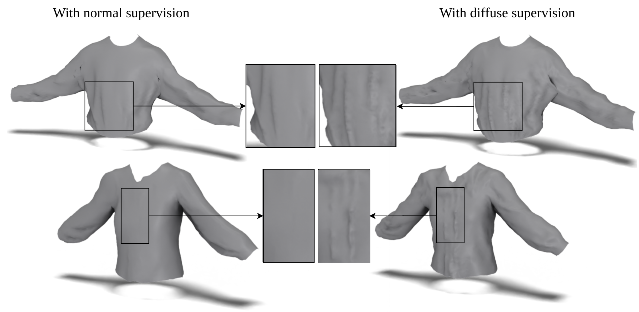

Figure 10. Ablative results comparing use of diffuse image supervision vs normal supervision.

are shown in Figure 7 where we demonstrate our method's ability to preserve high-frequency details superior to other SOTA methods. Overall, due to the implicit nature of representation, both SCARF [9] and REC-MV [33] fail to capture high-fidelity details in the garments' geometry. However, DGarments [26] addresses this limitation by predicting a per-vertex displacement on the explicit mesh. Nevertheless, their method is unable to model large deformations and hence struggles to handle loose garments effectively.

Texture Reconstruction : We present quantitative evaluations for novel view synthesis in Table 2, demonstrating that our method consistently outperforms the existing state-of-the-art across all visual evaluation metrics, including PSNR, SSIM [40], and LPIPS [45]. This highlights the high fidelity and perceptual quality of our approach. Additionally, the qualitative comparisons in Figure 8 further reinforce the effectiveness of our method where in terms of the visual quality of the appearance, our method yields sharp textural details in comparison to SCARF [9].

4.4. Ablation Studies

Effect of Adaptive Remshing : The effectiveness of our adaptive remeshing strategy is demonstrated in Table 1 (rows [3,4]). Although the CD & NC metrics show marginal quantitative degradation in case of without remeshing, we qualitatively demonstrate in Figure 9 that there is a significant drop in the fidelity of reconstructions (shown in row b), which is particularly leading to loss of complex folds and curved surfaces in comparison to reconstruction obtained with our full method (with remshing shown in row a). The remeshing process also effectively mitigates major artifacts by reducing the occurrence of larger triangles, as visible in the armpit region Figure 9 (see red circle). Furthermore, our remeshing strategy adaptively increases resolution in regions with higher geometric variation, enabling more precise capture of details such as folds, pockets, and other fine cloth structures (see Figure 9 square box), resulting in more accurate reconstructions.

Normals vs Diffuse Image : We observe that normals predicted (from [21]) in directions perpendicular to the viewing angles exhibit ambiguity. To address this limitation, we instead use diffuse images, which are basically the normals' components aligned with the viewing direction. Unlike standard normal maps, diffuse maps provide softer constraints, enabling improved generalization across frames. We provide empirical evaluation supporting the effectiveness of diffuse image supervision through ablation studies summarized in Table 1 (rows [3,5]). Our results consistently demonstrate that incorporating diffuse image supervision leads to improved performance compared to normal image supervision, further validating this design choice. A similar trend is observed in the qualitative comparisons illustrated in Figure 10, where the use of diffuse images results in improved geometric detail compared to normal image supervision.

5. Conclusion and Future Works

We propose a novel gradient-based deformation method to reconstruct dynamic textured garments from monocular video. We model both appearance and geometry and provide high-quality garment reconstruction. Our novel adaptive remeshing strategy further facilitates modeling high-frequency details and extremely loose garments. We demonstrate the superiority of our methods by showing improved qualitative and quantitative evaluations with SOTA methods. However, there is room for substantial improvement. One limitation of using a mesh representation instead of implicit functions is its susceptibility to self-intersection. Developing a more robust method to actively prevent self-intersections could significantly enhance results. Additionally, our deformations are not fully synchronized with environmental physics, sometimes leading to unrealistic movements. A more realistic solution would incorporate physics directly into the garment's deformation representation, beyond simply adding it as a loss term.

References

[1] Noam Aigerman, Kunal Gupta, Vladimir G. Kim, Siddhartha Chaudhuri, Jun Saito, and Thibault Groueix. Neural jacobian fields: learning intrinsic mappings of arbitrary meshes. *ACM Trans. Graph.*, 41(4), 2022. 2, 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8387–8397, 2018. CVPR Spotlight Paper. 5, 7

[3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Phorhum, 2022. 2

[4] Andrés Casado-Elvira, Marc Comino Trinidad, and Dan Casas. Pergamo: Personalized 3d garments from monocular video. In *Computer Graphics Forum*, pages 293–304. Wiley Online Library, 2022. 1, 2

[5] Lan Chen, Jie Yang, Hongbo Fu, Xiaoxu Meng, Weikai Chen, Bo Yang, and Lin Gao. Implicitpca: Implicitly-proxied parametric encoding for collision-aware garment reconstruction. *Graph. Models*, 129 (C), 2023. 2

[6] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people, 2021.

[7] Enric Corona, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Layernet: High-resolution semantic 3d reconstruction of clothed people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1257–1272, 2024. 2

[8] Marion Dunyach, David Vanderhaeghe, Loïc Barthe, and Mario Botsch. Adaptive remeshing for real-time mesh deformation. In *Eurographics 2013*. The Eurographics Association, 2013. 5

[9] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 1, 2, 6, 7, 8

[10] William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2, 3

[11] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 7

[12] 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, 2023. 2

[13] 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, 2024. 2

[14] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 1, 2

[15] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. pages 11026–11036, 2021. 2

[16] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion, 2024.

[17] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3090–3099, 2020. 2

[18] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, page 18–35, Berlin, Heidelberg, 2020. Springer-Verlag. 2, 7

[19] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 1, 2

[20] 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. 2

[21] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 7, 8

[22] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 2, 5

[23] Bruno Lévy, Sylvain Petitjean, Nicolas Ray, and Jérome Maillot. Least squares conformal maps for au-

tomatic texture atlas generation. *ACM Trans. Graph.*, 21(3):362–371, 2002. 5

[24] Ren Li, Benoît Guillard, Edoardo Remelli, and Pascal Fua. Dig: Draping implicit garment over the human body, 2022. 2

[25] Ren Li, Corentin Dumery, Benoît Guillard, and Pascal Fua. Garment recovery with shape and deformation priors, 2024. 2

[26] Xiongzheng Li, Jinsong Zhang, Yu-Kun Lai, Jingyu Yang, and Kun Li. High-quality animatable dynamic garment reconstruction from monocular videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 2, 3, 6, 7, 8

[27] Yifei Li, Hsiao-yu Chen, Egor Larionov, Nikolaos Sarafianos, Wojciech Matusik, and Tuur Stuyck. DiffAvatar: Simulation-ready garment optimization with differentiable simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[28] Lijuan Liu, Xiangyu Xu, Zhijie Lin, Jiabin Liang, and Shuicheng Yan. Towards garment sewing pattern reconstruction from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2023. 1, 2

[29] Luca De Luigi, Ren Li, Benoît Guillard, Mathieu Salzmann, and Pascal Fua. Drapenet: Garment generation and self-supervised draping, 2023. 2

[30] 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. *Communications of the ACM*, 65(1): 99–106, 2021. 1, 2

[31] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction innbsp;thenbsp;wild. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, page 184–200, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[32] 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, 2021. 2

[33] Lingteng Qiu, Guanying Chen, Jiapeng Zhou, Mutian Xu, Junle Wang, and Xiaoguang Han. Recmv: Reconstructing 3d dynamic cloth from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4637–4646, 2023. 1, 2, 5, 6, 7, 8

[34] Boxiang Rong, Artur Grigorev, Wenbo Wang, Michael J Black, Bernhard Thomaszewski, Christina Tsalicoglou, and Otmar Hilliges. Gaussian garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video. *arXiv preprint arXiv:2409.08189*, 2024. 1, 2

[35] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[36] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[37] Astitva Srivastava, Pranav Manu, Amit Raj, Varun Jampani, and Avinash Sharma. Wordrobe: Text-guided generation of textured 3d garments. In *European Conference on Computer Vision*, pages 458–475. Springer, 2025. 1

[38] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9065–9076, 2023. 4

[39] Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5, 6

[40] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 8

[41] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video, 2020. 2

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

[43] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. Econ: Explicit clothed humans optimized via normal integration, 2023. 2

[44] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20331–20341, 2024. 5

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 8

[46] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction, 2024. 2

[47] Yang Zheng, Qingqing Zhao, Guandao Yang, Wang Yifan, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, et al. Physavatar: Learning the physics of dressed 3d avatars from visual observations. *arXiv preprint arXiv:2404.04421*, 2024. 1

[48] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2020. 2

[49] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3845–3854, 2022. 2

[50] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023. 1