NGD: Neural Gradient Based Deformation for Monocular Garment Reconstruction

Soham Dasgupta Shanthika Naik Preet Savalia Sujay Kumar Ingle Avinash Sharma
Indian Institute of Technology Jodhpur
{sohamd, shanthikanaik, b22ai036, d23csa003, avinashsharma}@iitj.ac.in

1. Extended Related Works

Clothed Human Reconstruction: Clothed Human Reconstruction in itself is quite a challenging task, demanding generalization to different shapes and poses, tracking complex human motion while accommodating the garment's complex nature like non-rigid articulation, wrinkling, and their physical interaction with the parent body. Pifu [30] is one of the early works that used implicit representation to reconstruct clothed humans from a single image. [2, 11–13, 31, 37, 39, 41] further improves the reconstruction quality. While these methods provide good 3D reconstruction, they are temporally inconsistent when applied to videos. ICON [36] uses a two-stage pipeline first to obtain meshes from each frame of the video, and further employs Scanimate [32] to obtain temporally consistent animation of the clothed human meshes. However, learning from images always requires ground truth mesh data to learn from.

As compared to single images, monocular videos provide multi-view information in time, enabling richer reconstruction. Some methods [2, 35] use parametric body models like SMPL [24] as a base template and learn to deform them to match the clothed human body. Neural Body [27] utilizes the Neural Radiance Field for learning human representation and the underlying geometry extracted from the same representation. While neural rendering provides really good image reconstruction, it suffers from the limitations of the representation to extract good geometry. A line of work [9, 15] uses implicit representation which provides good reconstruction even for considerably loose garments. Reloo [10] further improves loose garment reconstruction by defining virtual bones on the garments and learning per-frame deformations. All the above-mentioned methods extract clothed humans as a single mesh and cannot extract garment mesh separately.

Image-based Garment Reconstruction: Several existing garment reconstruction methods [14, 26, 42] recover garments from monocular images. [14] deforms template mesh to represent garment image. More recent methods adopt implicit representation [3, 26] to represent the garments. DIG [19], Drapenet [25] learns to drape garment onto target SMPL body. ISP [20], Sewformer [23] utilize 2D sewing patterns to represent the garment which is stitched together to represent the 3D garment mesh. [21] further builds on ISP [20], to recover the garments from the image with improved accuracy. [19, 25] are garment draping methods that also reconstruct garments from images. [5] learns a latent garment space. [6] further improves high-frequency details for garment form images. All the aforementioned single-image reconstruction methods require supervised training on a large dataset.

2. Background

We assume any garment can be represented as a 2-manifold triangular surface mesh $\mathcal{S}$, in $\mathbb{R}^3$, with vertices $\mathcal{V}$ and triangles $\mathcal{T}$. Any deformation of a given triangular surface mesh $\mathcal{S}$ can be defined as a displacement function $\mathbf{d}$ which maps each point $\mathbf{p_i}$ to a displacement vector $\mathbf{d}(\mathbf{p}_i)$. A new surface $\mathcal{S}'$ can then be defined as follow:

$$\mathcal{S}' := \{\mathbf{p} + \mathbf{d}(\mathbf{p}_i) \mid \mathbf{p} \in \mathcal{S}\} \tag{1}$$

We further define this deformation as a piecewise linear displacement function $\mathbf{p} : \mathcal{S} \rightarrow \mathbb{R}^3, v_i \mapsto \mathbf{p_i}$, defined on the mesh surface. This piecewise linear function can be interpolated inside a triangle with vertices (v_i, v_j, v_k) as

$$\mathbf{p}(\mathbf{u}) = \mathbf{p_i}\mathbf{B_i}(\mathbf{u}) + \mathbf{p_j}\mathbf{B_j}(\mathbf{u}) + \mathbf{p_k}\mathbf{B_k}(\mathbf{u}) \tag{2}$$

where $\mathbf{u} = (u, v)$ is the 2D conformal map parameterization for the triangular surface mesh $\mathcal{S}$. $\mathbf{B}(\mathbf{u})$ is the orthonormal basis function to the tangent space of the triangle satisfying partition of unity $\mathbf{B_i}(\mathbf{u}) + \mathbf{B_j}(\mathbf{u}) + \mathbf{B_k}(\mathbf{u}) = 1$. The gradient of this function $\nabla\mathbf{p}(\mathbf{u})$ is defined as a uniform per triangle jacobians $\mathcal{J} \in \mathbb{R}^{3\times2}$ in the basis $\mathbf{B} = (\mathbf{B}_i, \mathbf{B}_j, \mathbf{B}_k) \in \mathbb{R}^{3\times2}$.

$$\begin{aligned}\mathcal{J} := \nabla\mathbf{p}(u) = (\mathbf{p}_j - \mathbf{p}_i)\nabla\mathbf{B_j}(\mathbf{u}) \\ + (\mathbf{p}_k - \mathbf{p}_i)\nabla\mathbf{B_k}(\mathbf{u})\end{aligned} \tag{3}$$

The gradient of $\nabla \mathbf{B_i}(\mathbf{u})$ can be defined as

$$\nabla \mathbf{B_i}(\mathbf{u}) = \frac{(v_k - v_j)^{\perp}}{2A_T} \tag{4}$$

where $\perp$ denotes the counter-clockwise rotation in the triangle space while A_T is the area of the triangle. The Equation 3 after substitution then gives rise to a constant Jacobian $\mathcal{J}$ for the given triangle in the surface mesh $\mathcal{S}$. The entire process can be described in matrix form for each triangle $t : t \in T$

$$\mathcal{J}_t = \mathbf{p} \nabla_t^T \mathbf{B} \tag{5}$$

where ∇_t^T can be considered as a discrete differential operator for the surface mesh $\mathcal{S}$.

For a piecewise deformation function $\mathbf{p} : \mathcal{S} \to \mathbb{R}^3, v_i \mapsto \mathbf{p_i}$, the gradient within the face is a constant $\mathbf{J} : \mathbf{J} \in \mathbb{R}^{3\times3}$. Note that here jacobian $\mathbf{J}$ is different $\mathcal{J}$ which was previously defined w.r.t the triangle basis $\mathbf{B}$., however, here it is defined w.r.t Euclidean basis $\mathbf{E}$. The gradient on each face $t : t \in T$ can now be defined as :

$$\mathbf{J} := \nabla \mathbf{p} = \begin{bmatrix} \nabla \mathbf{p_x} \\ \nabla \mathbf{p_y} \\ \nabla \mathbf{p_z} \end{bmatrix} \tag{6}$$

This jacobians $\mathbf{J}$ can then be modified by multiplying by matrix $\mathcal{M} : \mathcal{M} \in \mathbb{R}^{3\times3}$ which yields new set of face jacobians $\mathbf{J_t'} = \mathcal{M}_{\sqcup}\mathbf{J_t}$ for each face $t : t \in T$. The final deformation map $\mathbf{p}^*$ is obtained by making sure the gradients $\nabla \mathbf{p}^*$ is as close to the deformed jacobians $\mathbf{J}'$. In the continuous setting this becomes an energy minimization problem:

$$E(p) = \iint_{\mathcal{S}} ||\nabla \mathbf{p}(u,v) - \mathbf{g}(u,v)|| \mathbf{d}u\mathbf{d}v \tag{7}$$

where $\mathbf{g}$ is the continuous gradient field analogous to $\mathbf{J}'$. In a more discrete setting, the energy equations become:

$$\mathbf{p}^* = \min_{\mathbf{p}} \sum_{t \in T} |A_t| \, \|\nabla(\mathbf{p}) - \mathbf{J}_t'\|_2^2, \tag{8}$$

The solution can be obtained by solving the linear system formed by possion solved by knowing surface mesh cotangent Laplacian $\mathbf{L}$, and mesh mass matrix $\mathcal{A}$. The $\mathbf{J}'$ stacked to forms $\mathcal{M}$ which we use in our method to deform a base mesh.

For our use case, we handle the matrix $\mathcal{M}$ to dynamically deform the base mesh of the garment $\mathcal{S}$ in order to obtain the new garment mesh $\mathcal{S}'$. The primary benefit of using neural gradient-based deformation over other methods, such as the Laplacian-based deformation methods, is that it yields a smooth $\mathcal{C}^0$-continuous mesh. In contrast, $\mathcal{C}^1$-continuous methods, like those based on bi-laplacians, can cause over-smoothing, making them unsuitable for highly dynamic objects such as garments.

The primary benefit of using neural gradient-based deformation over other methods, such as the Laplacian-based deformation methods, is that it yields a smooth $\mathcal{C}^0$-continuous mesh. In contrast, $\mathcal{C}^1$-continuous methods, like those based on bi-laplacians, can cause over-smoothing, making them unsuitable for highly dynamic objects such as garments. The major benefit of this framework is that the predicted deformation field is triangulation agnostic while being in the gradient domain hence able to preserve highly accurate details.

3. Extended Methodology

3.1. Pose Encoding

Existing methods that condition the neural network with time parameter t tend to overfit each input view. While this improves reconstruction quality in the training view, the occluded areas lack high-frequency details, resulting in a smoothened surface. In contrast, conditioning on pose rather than time enables better feature alignment, as different viewpoints with the same pose can occur at various time instances, ensuring more consistent and coherent reconstructions. However directly using pose $\theta \in \mathbb{R}^d$ as a condition, where $d = 72$ for SMPL human body model leads to suboptimal training performance. Hence, instead of directly using the pose θ_t to the MLP, we use the PCA (Principle Component Analysis) for encoding the pose parameters as $\gamma(\theta_t)$.

For an input video with T frames, the overall pose is defined as $\theta \in \mathbb{R}^{T\times d}$. We find the PCA component of poses across all frames, $\theta_{PCA} \in \mathbb{R}^{T\times d}$. $\gamma(\theta_t) \in \mathbb{R}^k$ encodes the pose $\theta_t \in \mathbb{R}^d$ by projecting it onto the first k PCA components to get the reduced-dimension pose component. We use this PCA-projected pose component to condition f_φ. We find that projected pose conditioning helps in better generalization and prevents the neural network from overfitting to each frame. We set k as 4 for the geometric reconstruction module and 2 for the appearance module.

3.2. Gradient Based Adaptive Remeshing

We provide more details about our remeshing process over here. Post gradiant $\mathcal{G}(f_i)$ calculation on the faces f_i of the mesh M^B. We select the top quantile of faces $\mathcal{F}_\omega = \{f_i \mid \|\mathcal{G}(f_i)\| \geq \text{quantile}_\omega \left(\|\mathcal{G}(f_i)\|\right)\}$ for a percentile ω of triangle face. Subsequently, we prune all faces $\mathcal{F}_\delta = \{f_i \mid L(e_j) \geq \delta_{\text{length}}, \forall e_j \in \mathcal{E}(f_i)\}$ whose edge lengths fall below a certain threshold δ_{length}. Both the percentile α and edge length threshold δ_{length} are determined using a linearly decaying function $\alpha(t) = \alpha_0 + \gamma \cdot t$, $\Delta\delta_{\text{length}}(t) = \delta_0 - \Delta \cdot t$, ranging from high percentile and large edge length values to low percentile and small edge length values. This progressive strategy ensures stable remeshing that preserves crucial details without overloading the mesh with redundant ver-

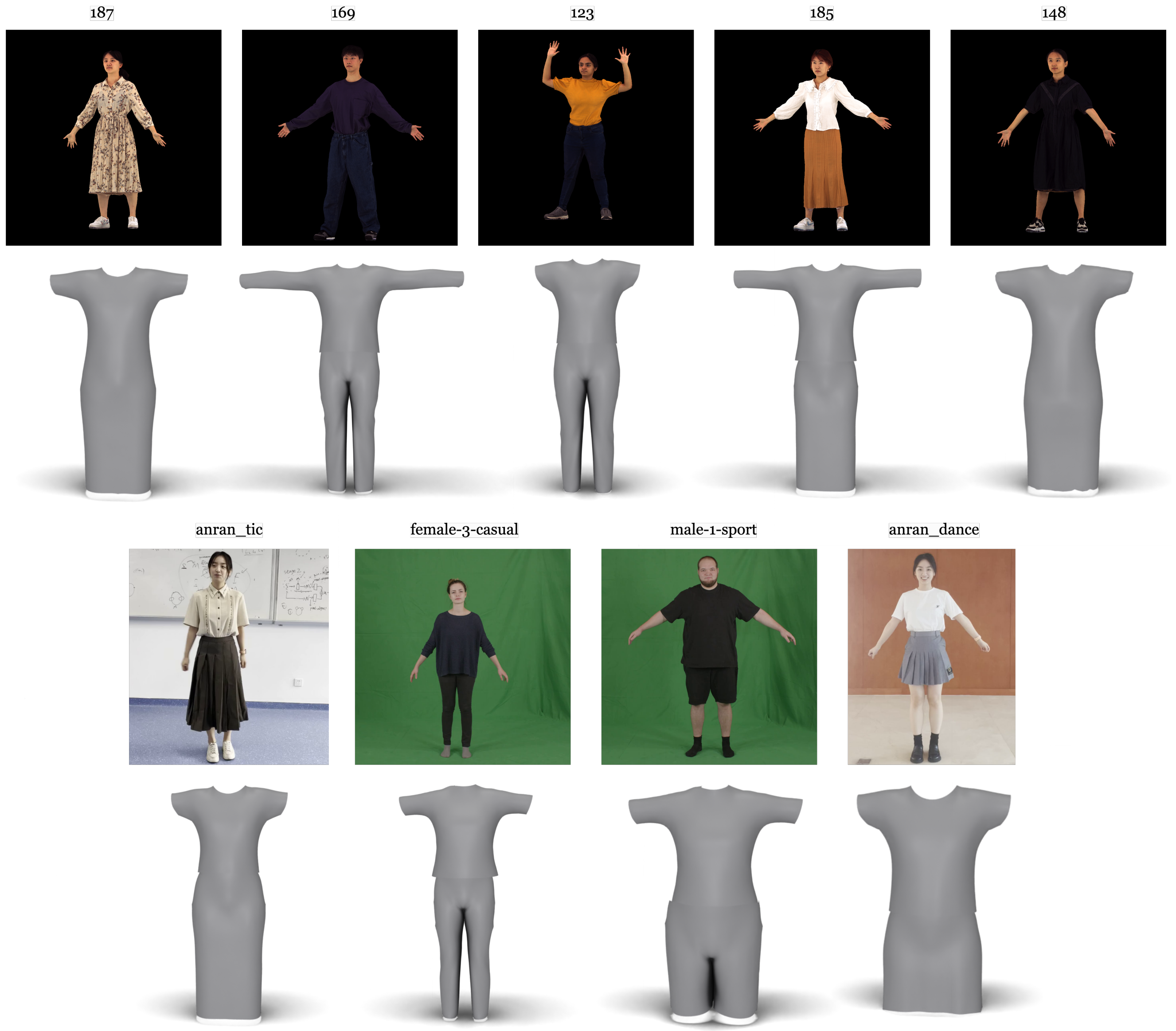

Figure 1. **Base Garment Meshes:** We present sample images from each video sequence with their corresponding base mesh employed in our experiments and evaluations.

tices. After the previous step, we obtain a face mask for each face $\mathcal{F}_\delta : \mathcal{F}_\delta \subseteq \mathbf{F}$ indicating the faces for remeshing. $SE = \{e_j \in \mathcal{E}(f_i) \mid f_i \in \mathcal{F}_\delta\}$.

The remeshing steps include two sequential operations - edge splitting and edge flipping as illustrated in Figure 2. We perform two sets of sequential operations on the set of selected edges E_s: edge splitting and edge flipping. In edge splitting, a selected edge is divided at its midpoint, creating two new triangles. Edge flipping adjusts the valence of each vertex, targeting a valence (number of neighbors at each vertex) of 6 for interior vertices and 4 for boundary vertices. The algorithm flips edges where the valence is too high; if flipping reduces the valence, the change is retained; otherwise, the edge is flipped back.

Post remeshing, the modified topology of the base mesh M_r^B after the re-meshing step requires the recomputation of all mesh attributes. The static deformation field J^S, being discrete and defined on previous mesh topology, contrasts to the continuous nature of the dynamic field J_t^D, requires the recomputation of the discrete Jacobians field J^S. Since Jacobians J^S is still essentially an extrinsic field and doesn't depend on the triangulation of the mesh, we can interpolate the new set Jacobians $\tilde{J^S}$ from the previous set J^S using K nearest neighbors interpolation. Similarly, for new op-

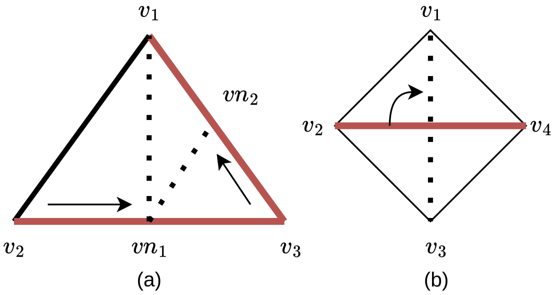

Figure 2. **Remeshing Operations:** (a) Edge split (colored in red) into new vertices vn_1 and vn_2. (b) Edge flipping

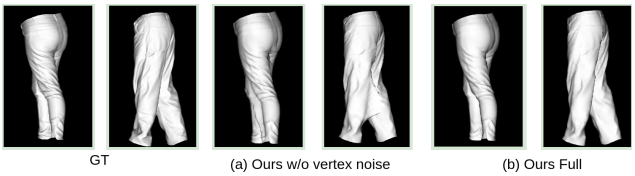

Figure 3. **Advantage of vertex loss:** We can successfully prevent local minimums by preventing premature convergence

timization parameters, including the first and second moments m_1^r and m_2^r for the Adam optimizer [17], are also interpolated after each remeshing step to maintain consistent training and prevent jerk in the optimization. Finally, the new skinning weights W^r are also computed by interpolating the previous skinning weights on the mesh surface. Specifically, the newly interpolated m_1^r, m_2^r, J^r and W^r at each face center is computed as follows:

$$\omega_k = \frac{\frac{1}{d_k+\epsilon}}{\sum_{k=1}^{K} \frac{1}{d_k+\epsilon}} \tag{9}$$

$$\zeta_j^r = \sum_{k=1}^{K} \zeta_j \cdot \omega_k; \quad \zeta \in (m_1, m_2, J^S, W), \tag{10}$$

where ζ_j^r represents post-remeshed mesh attributes and d_k represents the distance between the original and changed face centers for the k^{th} nearest neighbor and ϵ is added for numerical stability.

3.3. Dealing with Local Minima

Vertex Noise: One common challenge in reconstructing clothing, such as shirts and pants, is that the optimization process often becomes trapped in local minima as shown in Figure 3. In an effort to minimize local rendering losses, the global geometry is inadvertently distorted, leading to unrealistic artifacts. To address this, we introduce a novel exponentially decaying noise applied to the vertices v_i of the final skinned mesh M_t^P at each iteration:

$$\mathcal{X}(t) = \mathcal{N}(0,1) \cdot \left(1 - \frac{e}{\eta}\right) \tag{11}$$

$$\tilde{v}_i = v_i + \mu \cdot \mathcal{X}(t) \tag{12}$$

where $\mathcal{N}(0,1)$ is a standard Gaussian distribution, μ is the noise scaling parameter and e and η represent the current and total iterations, respectively. This noise encourages the model to prioritize global geometry in the initial stages, preventing early overfitting to local details. As training progresses, the noise and learning rate both gradually reduce, allowing the model to refine local geometric features without distorting the overall structure. Importantly, this smooth noise application introduces no additional computational overhead while significantly improving the reconstruction quality of loose garments.

3.4. Mask Loss

A common challenge in garment reconstruction is self-occlusion, where parts of the garment, like the hands, occlude each other in the image. This occlusion also appears in the mask generated from the input image, but the corresponding rendered meshes lack such occlusions, leading to inconsistencies during training and causing artifacts. REC-MV [28] addresses these self-occlusion inconsistencies by utilizing *feature lines*, however, obtaining accurate feature lines across frames is expensive and labor-intensive. Instead, we employ part segmentation from Sapiens [16] to obtain hand masks I_{occ}^s, which help in identifying significant self-occlusions (as hands are mostly the primary cause). These occlusion masks are used to modify the mask loss and reduce occlusion-induced artifacts. The mask loss is formally defined as:

$$\mathcal{L}_{\text{mask}} = ||(I_{\text{pred}}^s \odot (1.0 - I_{\text{occ}}^s)), I_{\text{gt}}^s||_2^2 \tag{13}$$

3.5. Data Preprocessing

We utilize existing pre-trained vision models to obtain reliable priors. We extract SMPL shape $\beta_t \in \mathbb{R}^{10}$ and per-frame pose $\theta_t \in \mathbb{R}^{N\times 72}$ parameters as well as per-frame camera estimation $\pi_t = (R, T) : R \in \mathbb{R}^{N\times 3\times 3}, T \in \mathbb{R}^3$ using 4DHumans [8], an SMPL fitting method for monocular video. Per-frame normal map, depth map, and part-segmentation are recovered using a pre-trained human foundation model Sapiens [16], which serves as pseudo-ground truth during optimization. Finally, a t-pose base garment M^B is obtained from the first frame using BCNet [14]. Our method is robust to the quality of these base meshes (details in Suppl.). We obtained the garment skinning weight $W : W \in \mathbb{R}^{N\times 24}$ interpolated from posed body skinning weight $\mathcal{W}$. The interpolation is performed using Gaussian Radial Functions (GRBF) [4] with Gaussian Kernel

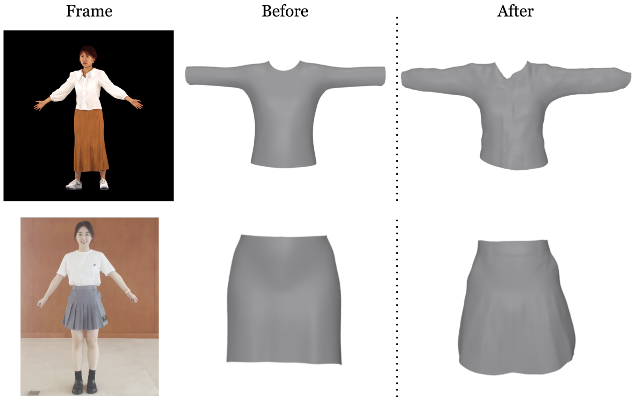

Figure 4. **Global garment shape:** We can significantly transform our garment from the initial base mesh.

$\Omega(r) = e^{-\frac{r^2}{km_i^2}}$ where m_i is the Euclidean distance between a garment vertex $v_i \in M^B$ and its nearest body vertex.

Base Garment Meshes : We visualize all the base garments used for our evaluations Figure 1. Most of the templates are generated from BCNet [14] with few exceptions which are derived from existing garment assets. Please note that we utilize the base mesh with only shape-specific deformations provided by BCNet. Furthermore, as illustrated in Figure 4, our global shape deformation effectively morphs the base mesh, substantially altering its structure to approximate the desired shape.

4. Experiments

4.1. Dataset

4DDress-Mono : Dress4D [33] is a real-world 4D dataset of textured clothed humans in diverse motion sequences. It provides dynamic meshes with vertex-level semantic annotations for garments and the human body, combined with fitted SMPL(-X) models. However, the default camera setup in Dress4D does not capture a full 360° view of the garment from a single camera, making it challenging to record details across frames and register them onto the garment. To address this, we create a modified dataset by initializing a virtual camera and generating a 360° video of the subjects, enabling a more comprehensive benchmark for our model against baseline methods.

Additional Datasets: We evaluate our method on two sequences from the PeopleSnapshot dataset [1] and two from REC-MV [28] video recordings. While PeopleSnapshot features standard clothing with a close fit and simple motions, REC-MV includes sequences with loose garments, such as skirts, which are crucial for assessing the robustness of our approach under diverse conditions.

Evaluation Protocol : We provide qualitative comparisons for the 4DDress-Mono dataset and evaluate surface reconstruction using Chamfer Distance $\mathcal{L}_2$ and normal consistency (NC). For texture recovery, we assess novel view synthesis quality using PSNR, SSIM [34], and LPIPS [38]. Additionally, we present qualitative comparisons for both surface reconstruction and novel view synthesis on this dataset. For additional datasets such as PeopleSnapshot [1] and Rec-MV [28], we report only qualitative comparisons of mesh surface quality.

Choice of SOTAs: Given the limited number of works on monocular garment reconstruction, we rely on the following state-of-the-art methods: SCARF [7], Rec-MV [28], and DGarments [22].or the 4D-Dress dataset, we utilized the ground truth SMPL mesh as the base for reconstruction in SCARF [7], DGarments [22], and our proposed method. This ensured consistency in surface reconstruction evaluations across all approaches. However, for sequence 148, we were unable to obtain a plausible reconstruction using SCARF, and thus, we omit its results for this sequence. Additionally, the texture generation code for DGarments [22] was not provided, limiting our ability to evaluate its appearance module. For Rec-MV [28], the absence of data preprocessing code—specifically, the feature line estimation required for extracting garment structure—prevented us from conducting direct comparisons on our custom dataset. Nevertheless, we present qualitative results on additional datasets [1, 28], as the preprocessed files for these datasets were available. For these additional datasets, we selected a subset where the training sequences consisted of a single circular motion to ensure consistency in evaluation.

4.2. Additional Experiments

4.3. Comparisons

We provide more qualitative comparisons for both surface reconstruction and novel view synthesis comparison experiments of our method with the SOTAs [7, 22, 28] given in Figure 9, Figure 10 and Figure 11. Similar to the previously shown results, our method consistently achieves high-quality reconstructions, capturing finer details with minimal artifacts. By decoupling geometry and appearance reconstruction, our approach enables more accurate and realistic reconstruction of both.

We further present qualitative results across multiple frames to demonstrate the dynamic reconstruction capability of our method, as shown in Figure 5. Our reconstructed results from the proposed method seem to remain consistent across time frames. The REC-MV authors provided a preprocessed, sampled version of their dataset and the PeopleSnapshot dataset, but neither includes groundtruth novel view data, making direct comparisons for novel view synthesis infeasible. However, we have included a qualitative comparison of novel view generation (sans available ground truth view) in Figure 6 using their provided results. The

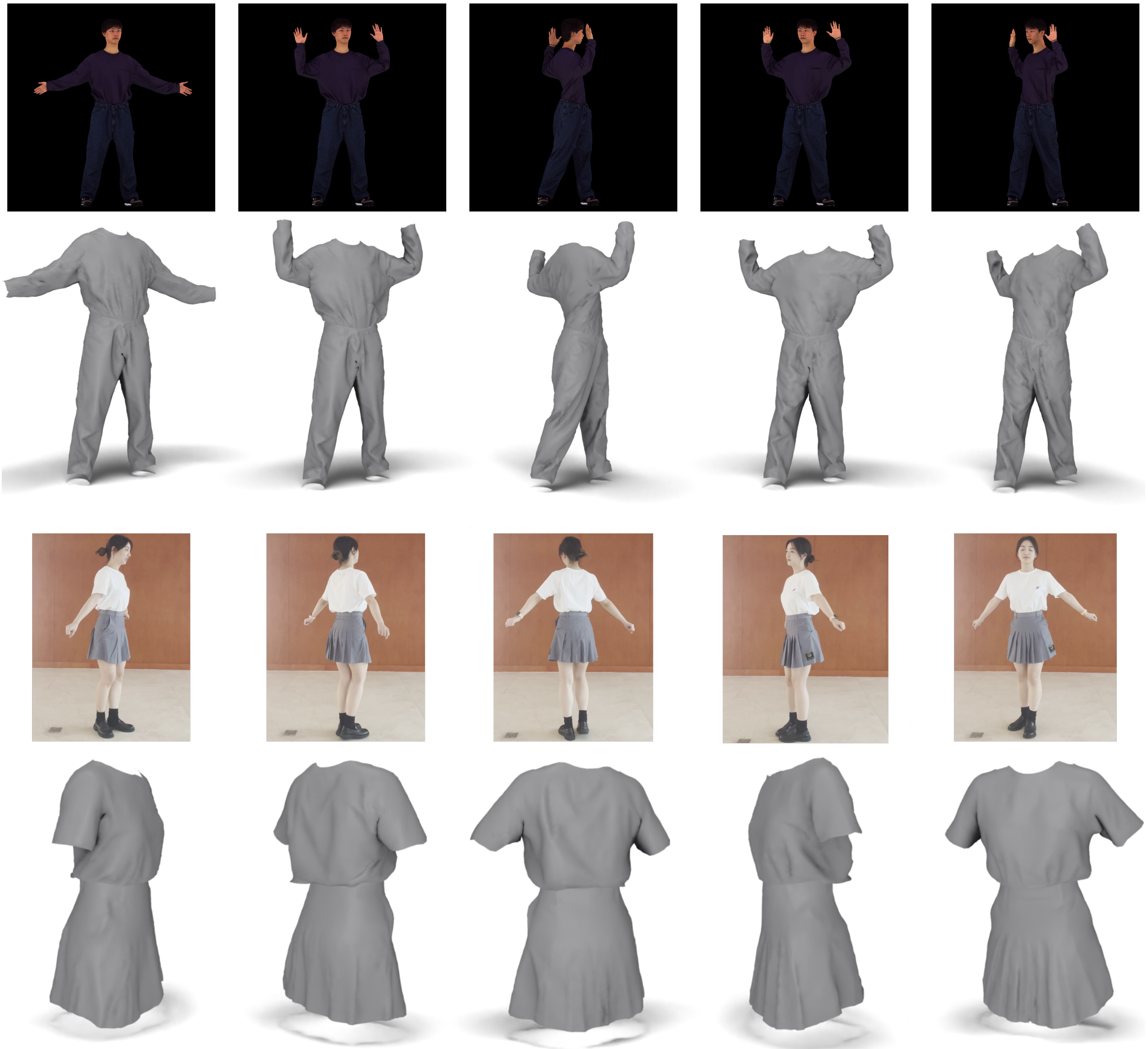

Figure 5. **Dynamic garment reconstruction results of our method** for one sequence from 4D-Dress [33] and one sequence from Rec-MV [28] dataset

aforementioned figure shows that our method yields superior texture reconstruction in novel views, in comparison to REC-MV.

4.4. Extended Ablative Studies

Effect of Shape Regularizer: The Jacobian regularizer plays a crucial role in constraining triangle deformations, ensuring that they do not deviate significantly from their original shapes. By preserving the global geometric structure, the regularizer prevents excessive divergence from the base mesh, which could otherwise introduce noticeable artifacts in under-constrained novel views. To illustrate these phenomena, we provide qualitative comparisons in Figure 4.4 to demonstrate the effectiveness of $\mathcal{L}_{\text{reg}}$. In the absence of this regularization term, the reconstruction deviates significantly from the original shape, leading to substantial artifacts. Without the regularizer, our proposed method performs poorly on tighter clothing, such as shirts, where shape preservation is critical.

Effect of Vertex Noise : Cloth reconstruction is a highly underconstrained problem, particularly in monocular settings. This challenge is further exacerbated when using de-

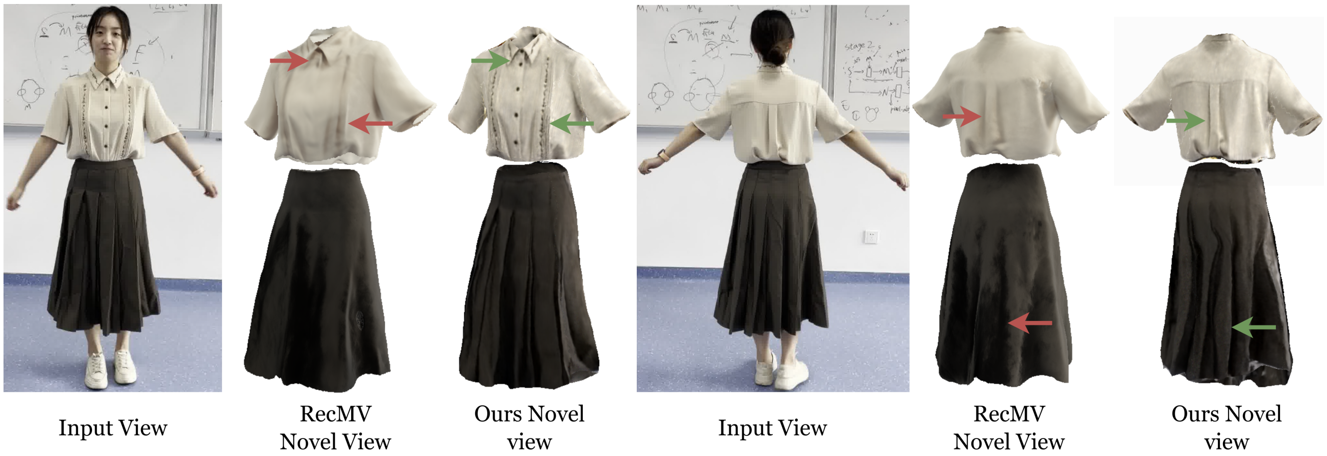

Figure 6. The novel-view synthesis results of REC-MV and Ours.

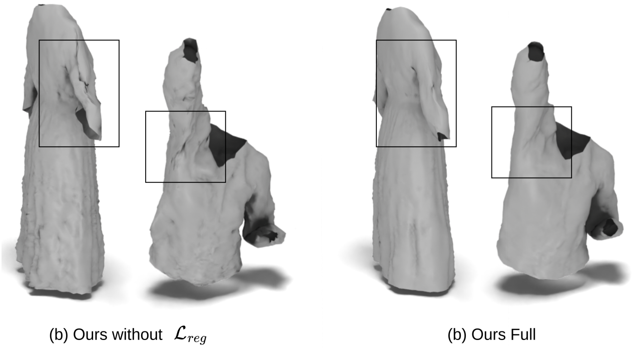

Figure 7. **Importance of Regularization:** Regularization helps maintain the template structure by preventing excessive deformations, which could otherwise lead to unrealistic results.

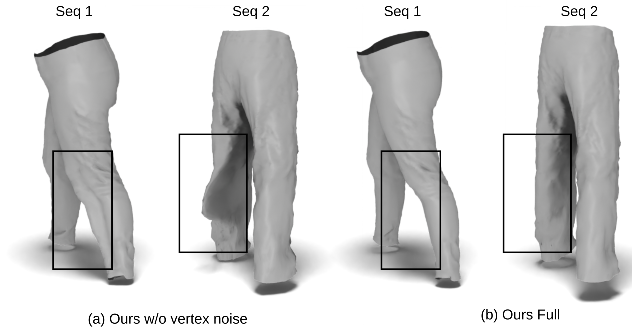

Figure 8. **Importance of Vertex Noise:** Without vertex noise, our method may converge to local minima, leading to unnatural folds of the reconstructed garments.

ferred shading-based differentiable rendering engines like nvdiffrast [18]. During the reconstruction of fine details, such as small folds, the method often misaligns garments and converges prematurely to local minima, from which it fails to recover. As a result, it attempts to reconstruct garment boundaries as folds in an undesirable manner, as shown in Figure 3. This issue is more pronounced when using L1 loss instead of Huber loss, though even with the latter, it is not entirely avoided. To address this, we introduce an exponentially decaying vertex noise strategy, particularly for garments prone to local minima (e.g., pants). This technique prevents premature convergence during early optimization stages, promoting a more global reconstruction. As illustrated in Figure 8, vertex noise significantly enhances reconstruction quality, with quantitative results provided in Table 1.

Table 1. comparison of geometry reconstruction on Dress4D [33].

Method	Chamfer Distance $\mathcal{L}_2 \downarrow$	NC $\uparrow$
Ours w/o Vertex Noise	0.091	0.910
Ours Full	**0.088**	**0.912**

5. Discussion and Limitations

Temporal Consistency : Although our method demonstrates reasonably good temporal consistency across frames due to dynamic prediction from a single MLP, it does not guarantee full temporal coherence.

Physics Based Constrains : Monocular reconstruction is a highly underconstrained problem. Novel views outside the visible region are particularly ill-posed, constrained only by weak regularization. As a result, reconstructions in these regions can be physically implausible. Introducing physics-based constraints from simulators [29, 40] can be especially useful in achieving more realistic deformations in these unconstrained regions.

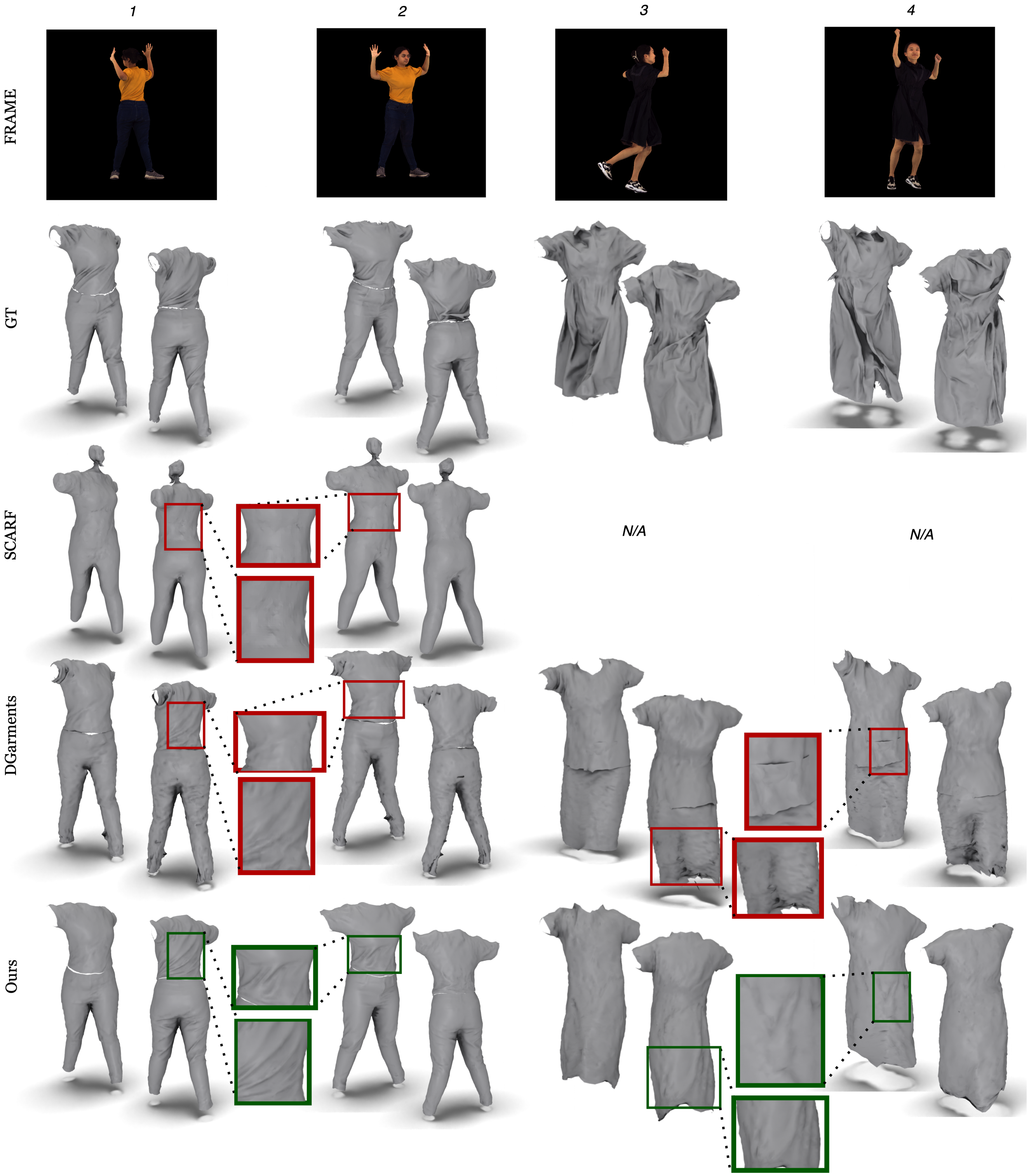

Figure 9. **Qualitative comparison on 4D-Dress [33]**

References

[1] 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 Pat-*

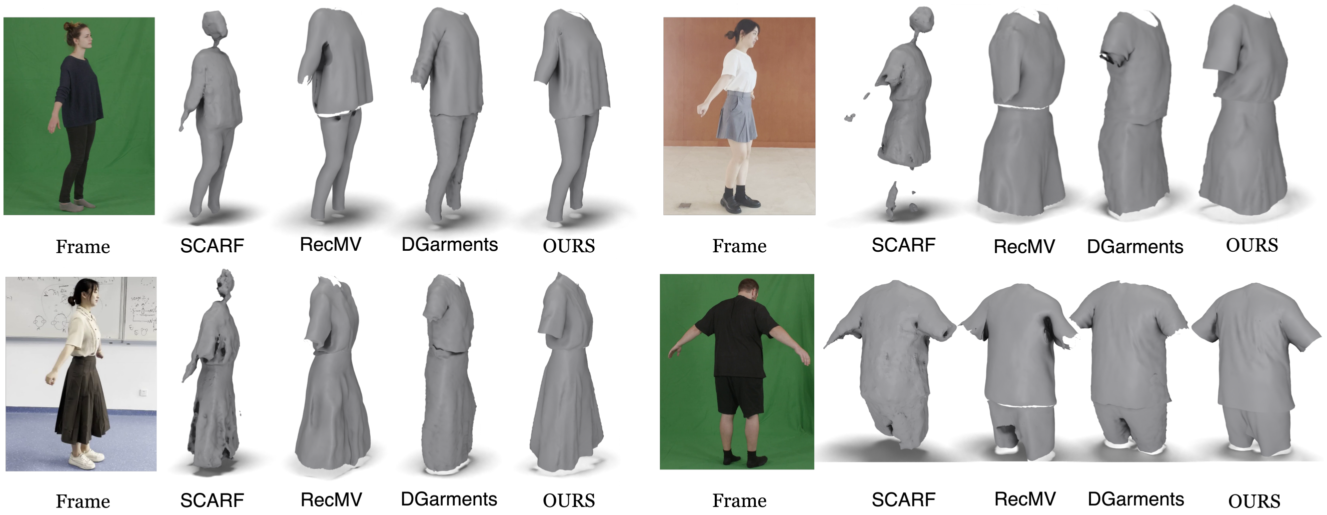

Figure 10. **Qualitative comparisons on People Snapshot [1] dataset**

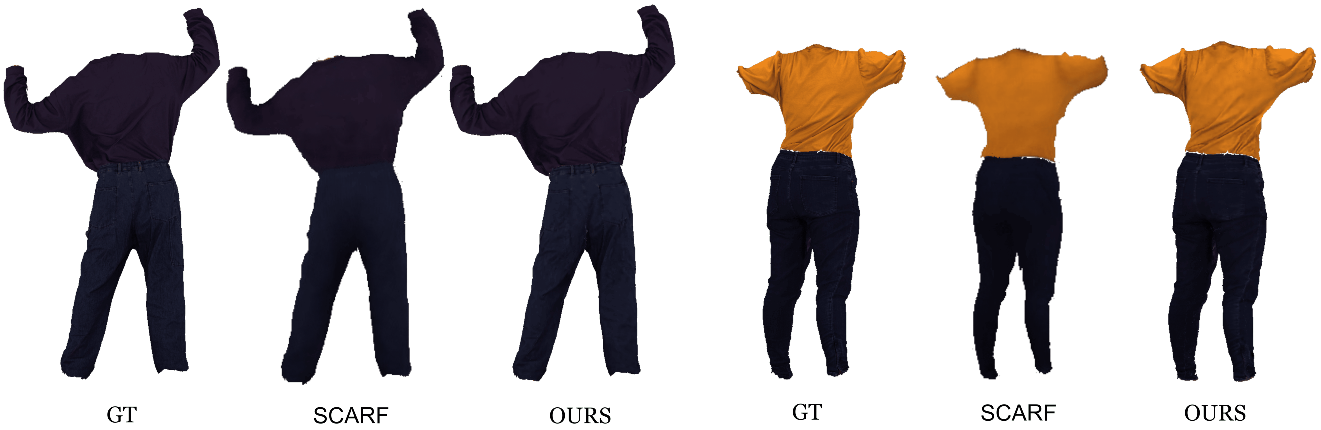

Figure 11. **Qualitative comparison novel view synthesis**

tern Recognition (CVPR), pages 8387–8397, 2018. CVPR Spotlight Paper. 5, 9

[2] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Phorhum, 2022. 1

[3] 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. 1

[4] Ruochen Chen, Shaifali Parashar, and Liming Chen. Gaps: Geometry-aware, physics-based, self-supervised neural garment draping. In *International Conference on 3D Vision (3DV)*, 2024. 4

[5] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people, 2021. 1

[6] 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. 1

[7] 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. 5

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

[9] 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. 1

[10] 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. 1

[11] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. pages 11026–11036, 2021. 1

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

[13] 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. 1

[14] 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. 1, 4, 5

[15] 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

[16] 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. 4

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

[18] 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. 7

[19] Ren Li, Benoît Guillard, Edoardo Remelli, and Pascal Fua. Dig: Draping implicit garment over the human body, 2022. 1

[20] Ren Li, Benoît Guillard, and Pascal Fua. Isp: Multi-layered garment draping with implicit sewing patterns, 2023. 1

[21] Ren Li, Corentin Dumery, Benoît Guillard, and Pascal Fua. Garment recovery with shape and deformation priors, 2024. 1

[22] 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. 5

[23] 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

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 1

[25] Luca De Luigi, Ren Li, Benoît Guillard, Mathieu Salzmann, and Pascal Fua. Drapenet: Garment generation and self-supervised draping, 2023. 1

[26] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction in the 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. 1

[27] 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. 1

[28] 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. 4, 5, 6

[29] 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. 7

[30] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 1

[31] 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. 1

[32] 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. 1

[33] 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, 7, 8

[34] 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. 5

[35] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video, 2020. 1

[36] 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. 1

[37] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. Econ: Explicit clothed humans optimized via normal integration, 2023. 1

[38] 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. 5

[39] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction, 2024. 1

[40] 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. 7

[41] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2020. 1

[42] 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. 1