Supplementary Material for
"REC-MV: REconstructing 3D Dynamic Cloth from Monocular Videos"

Lingteng Qiu[1*] Guanying Chen[1,2*] Jiapeng Zhou[1] Mutian Xu[1]
Junle Wang[3] Xiaoguang Han[1,2†]
[1]SSE, CUHKSZ [2]FNii, CUHKSZ [3]Tencent

Contents

1. Supplementary Video

Please check the attached video for dynamic garment reconstruction results of our methods.

2. More Details for the Method

2.1. Explicit Garment and Surface Templates

The adopted garment template covers ten common clothes categories. It includes the long/short/no sleeve upper clothing, and tube dress categories (see Fig. S1 (a)). Note that the sleeve upper clothing templates are shared for the dress templates (*i.e.*, long/short/no sleeve dress). Figure S1 (b) shows the templates that are designed for long pants, short pants, and skirts. For surface templates, the bottom surface template is the same as the garment template. For the upper surface template, we include the head and hand parts in the garment template, as shown in Fig. S1 (c). Our design of the surface template can improve the reconstruction results (see Sec. 5.3 of this supplementary material for examples).

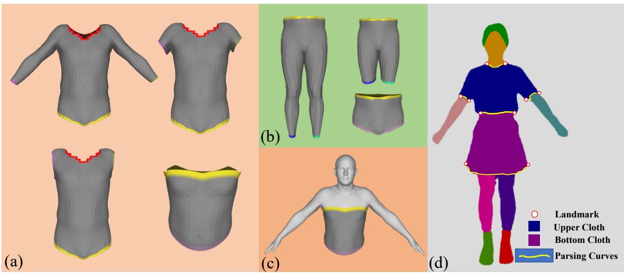

Figure S1. Illustration of the garment templates, surface template, and the 2D parsing curve diagram. (a) The upper clothing templates. (b) The bottom clothing templates. (c) The surface template with tube dress type. (d) Parsing 2D visible curve from semantic mask and landmarks.

2.2. Parsing Visible Curve from Segmentic Mask

In our work, we need to automatically parse 2D garment curves. However, there is no existing method for 2D garment curves estimation from monocular video. We observe that visible 2D feature curve can be considered as the boundary of clothing mask. As shown in Fig. S1 (d), the 2D feature curves can be parsed from the shortest path on clothed boundary according to the predicted garment landmark points. Specifically, we train both the HigherHRNet [2] and Semantic Network [8] using DeepFashion2 dataset [4] to predict the garment landmarks and the clothed boundaries.

2.3. Differentiable Surface Rendering

we utilize an MLP f_c with learnable parameters ψ to model the color of surface points in the canonical space. Given a camera ray $\mathbf{v}$ with the camera center $\mathbf{c}$, we first use differentiable non-rigid ray-casting method [6] to find the intersection points $\mathbf{p}$ on $S(\eta)$ by solving

$$\mathbf{p} = \arg\min_{\hat{\mathbf{p}}} \lambda |f(\hat{\mathbf{p}})| + \frac{||(\Phi(\hat{\mathbf{p}}) - \mathbf{c}) \times \mathbf{v}||}{\Phi(\hat{\mathbf{p}}) - \mathbf{c}}. \quad (1)$$

[*] Equal contribution.
[†] Corresponding author: hanxiaoguang@cuhk.edu.cn.

To make the above equation differentiable, we solve some normal equations to obtain their differentiable formulation. Specifically, the surface points $\mathbf{p}$ is constrained by both implicit surface function f and the camera ray:

$$\begin{aligned} f(\mathbf{p}) &\equiv 0 \\ (\Phi(\mathbf{p}) - \mathbf{c}) \times \mathbf{v} &\equiv 0 \end{aligned} \tag{2}$$

where Φ is deformation field, $\mathbf{c}$ is the camera center and $\mathbf{v}$ is a camera ray.

For learnable weights η of implicit surface f, we differentiate these two equations w.r.t η to obtain:

$$\begin{aligned} \frac{\partial f^T}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \eta} &= -\frac{\partial f}{\partial \eta} \\ [\mathbf{v}]_\times \frac{\partial \Phi(\mathbf{p})}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \eta} &= 0 \end{aligned} \tag{3}$$

In terms of parameters ϕ of deformation field Φ, we have the following constrain:

$$\begin{aligned} \frac{\partial f^T}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \phi} &= 0 \\ [\mathbf{v}]_\times \frac{\partial \Phi(\mathbf{p})}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \phi} &= -[\mathbf{v}]_\times \frac{\partial \Phi(\mathbf{p})}{\partial \phi} \end{aligned} \tag{4}$$

where $[\mathbf{v}]_\times$ denotes as $\mathbf{v}$'s cross product matrix. Then, we could get differentiable formula $\frac{\partial \mathbf{p}}{\partial \eta}$ and $\frac{\partial \mathbf{p}}{\partial \phi}$ from the above 4×3 linear equations.

2.4. Optimizing Sequences with Large Motion

Modeling garment surface deformations in a video sequence with large motions is a challenging problem, especially for dresses and skirts. To better reconstruct dynamic garment surfaces with large motion, we first reconstruct the garment surface in the canonical space from a sequence depicted self-rotated human, we then freeze the weights of the implicit surface η, and only optimize the latent code of large pose frame and deformation field.

2.5. More Details for Loss functions

Normal Regularization Loss. To further refine geometry, we first employ pix2pixHD to predict clothed human normal map $\{\hat{N}_t | t = 1, ..., N\}$ [12]. Then, following Jiang et al. [6], given $\mathbf{p}$ in canonical view and its corresponding points $\mathbf{q}$ in camera-view, we obtain $\mathbf{n_p} = \nabla f(\mathbf{p}; \eta)$ and its ground truth by transforming the normal of corresponding point $\hat{N}\mathbf{q}$ to the canonical space, which can be computed by $J_{\mathbf{q}}(\mathbf{p})^T \hat{N}_{\mathbf{q}}$. Hence, The surface normal loss can be computed by:

$$\mathcal{L}_{norm} = \frac{1}{|\mathcal{R}|} \sum_{p \in \mathcal{R}} \lambda_p ||\mathbf{n_p} - J_{\mathbf{q}}(\mathbf{p})^T \hat{N}_{\mathbf{q}}||_2, \tag{5}$$

where λ_p is the weight defined by the cosine of angle between $\mathbf{n_p}$ and corresponding view direction [6].

Rigidity Loss. To avoid distortion of non-rigid transformation, following Park *et al.* [11], a rigid loss $\mathcal{L}_{arap}$ is computed to constrain the non-rigid deformation D_g that should be as rigid as possible:

$$\mathcal{L}_{arap} = \sum_{g=1}^{N_g} \frac{1}{|S_g|} \sum_{\mathbf{p} \in S_g} \rho(||log\Sigma_p^g||_F), \tag{6}$$

where Σ_p^g is the singular value matrix of the Jacobian of $\mathcal{D}_g$ on $\mathbf{p}$ and ρ is the robust funtion [3].

Eikonal Loss. The Eiknoal loss of IGR [5] is adopted to regularize the gradient of f_g, make this function being sign distance function:

$$\mathcal{L}_{eik} = \sum_{g=1}^{N_g} \frac{1}{S_g} \sum_{\mathbf{p} \in S_g} (||\mathbf{n_p}||_2 - 1)^2. \tag{7}$$

2.6. Garment Extraction from Implicit Surface

Template Deformation. We employ N_T garment templates $\{\mathbf{T}_i | i = 1, \ldots, N_T\}$ as DeepFashion3D [14]. After obtaining the 3D feature curve sets that belong to a garment template $\mathbf{T}_i$, The handle-based Laplacian deformation [13] is utilized to obtain $\bar{\mathbf{T}}_i$ from $\mathbf{T}_i$ so that its feature curves fit the optimized curves.

Surface Probing in Implicit Function. Given the $\bar{\mathbf{T}}_i$ that belongs to the current implicit surface, we first employ $f(\mathbf{p}; \eta)$ to generate garment surface $\mathbf{T_s}$ via marching cube algorithm [10]. Next, we adopt an adaptive non-rigid ICP to transfer high-frequency details. Specifically, we only optimize the valid corresponding points whose angle of normal directions is less than a preset threshold δ (δ is set as $60°$). The adaptive non-rigid ICP helps to remove erroneous correspondence and produce our final reconstruction $\mathbf{T}'$. Then we can utilize the deform field Φ to warp $\mathbf{T}'$ to camera view space based on the pose parameters and the corresponding per-frame latent code $\mathbf{h}_i$.

3. Implementation Details

2D Visible Curve Parsing. For the 2D visible curve parsing network, we train HigherHRNet [2] to predict two garment landmark on DeepFashion2 dataset [4]. With weights pretrained on COCO [9], we train the model for 13 epochs, where we utilize Adam optimizer [7] and set its learning rate to 0.001. It costs 27 hours to train it on four GTX 3090 GPUs.

Curve and Surface Optimization. We jointly optimize both the explicit curves and implicit surface for 200 epochs from scratch. The Adam optimizer [7] is used to optimize our implicit and explicit parameters. For implicit surface

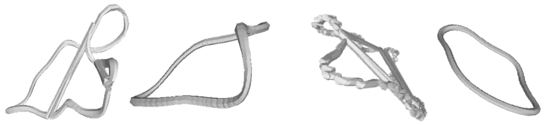

Figure S2. Ablation study for the intersection-free curve deformation. Inter-free is short for intersection-free deformation.

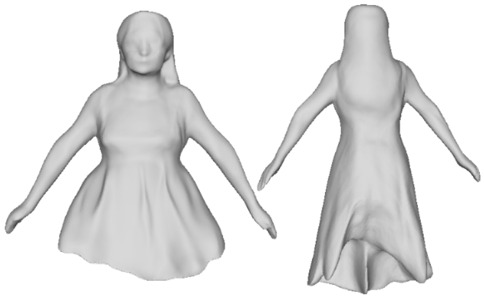

Figure S3. Implicit surface reconstruction results of the model without the curve-aware surface initialization.

optimization, the learning rate is set as 0.0001 with a decay rate of 0.3 every 50 epochs. The learning rate of the explicit curve optimization is set as 0.001. It takes about 32 hours per 300 frame to train our method on one GTX 3090 GPU.

Large Pose Optimization. With respect to capture the sequence with large pose, we only optimize latent code and deformation weights with a learning rate set as 0.0001. It spends approximate 24 hours fitting a video containing 260 frames on one GTX 3090 GPU.

4. More Details for the Dataset

For the diversity of garment category, we utilize the PeopleSnapshot [1] dataset and seven real sequences collected by ourselves for qualitative evaluation.

Specifically, the sequence we use in PeopleSnapshot are *female-3-casual*, *female-3-sport*, *female-4-casual*, *female-6-plaza*, *male-1-casual*, *male-2-casual*, *male-2-outdoor*, *male-4-casual*, *male-5-outdoor*, and *male-9-plaza*.

5. More Results for Ablation Study

5.1. Intersection-free Curve Deformation

Figure S2 shows the ablation study for intersection-free curve deformation. If directly regressing offsets of points in the curve, we observe that the curve is prone to self-intersection and destroy its original order regardless of including the edge smooth loss. Note that the computation of curve-guided implicit consistency loss $\mathcal{L}_{ccons}$ requires the curve points to maintain their order, which further demonstrates the strengths of our intersection-free curve deformation method.

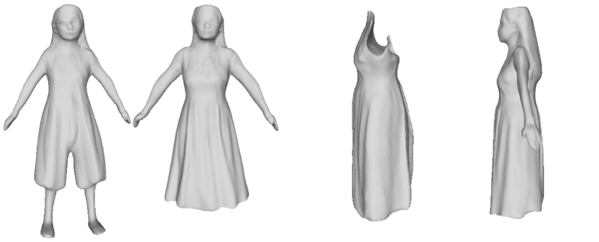
(a) Body Temp. vs. Portrait Temp.
(b) Garment Mask vs. Portrait Mask

Figure S4. Ablation study for surface templates initialization. (a) Comparison of results of models initialized with body template and portrait template. (b) Comparison of results with garment mask and portrait mask supervision.

5.2. Curve-aware Surface Initialization

Figure S3 illustrates that implicit surface is prone to produce collapsed surface without curve-aware initialization. This is because if the distance between initial surface and target surface is very large, the non-rigid deformation field has to learn these long-distance offsets. However, it is hard for MLP to fit long-distance deformation without ground-truth surface supervision.

5.3. Surface Templates Initialization

We conduct the experiment to explain why we need different surfaces templates for initialization. As Fig. S4 (a) illustrates, if we only use the whole-body as the surface template to initialize garment implicit surface, it hardly learns the shape of hemlines. Hence, we use three different surface templates to initialize implicit surface.

Note that we employ the portrait (*i.e.*, including the garment, head and hand parts) as our upper cloth template instead of upper-cloth garment template. The reason is that only using garment mask to supervise implicit surface leads to collapse surface artifacts, as shown in Fig. S4 (b). Therefore we adopt the portrait template to initialize upper cloth implicit surface.

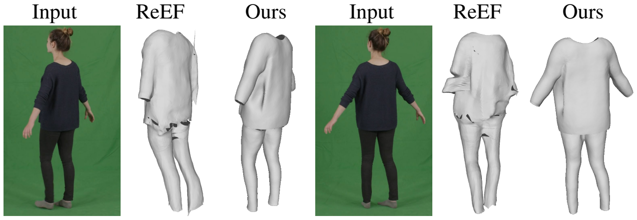

Figure S5. Comparison between ReEF [15] and our method on two frames in a video sequence.

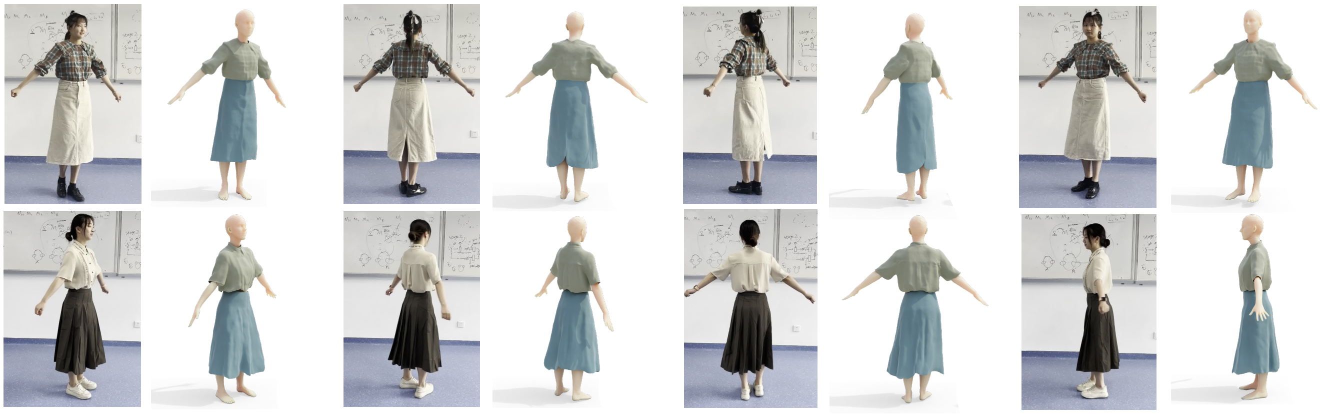

Figure S6. More dynamic reconstruction results on our method. Each row shows the reconstruction of four frames in a monocular video.

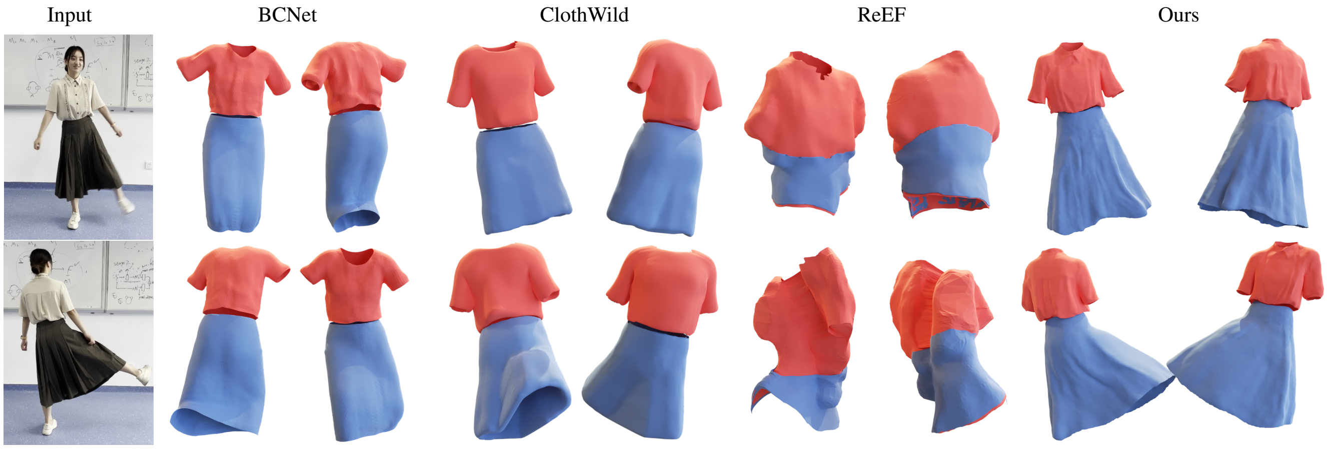

Figure S7. Reconstruction results on a large pose sequence. Each row shows the key frame of the video and results of different methods.

5.4. Temporal Consistency

As our method represents the garment in the canonical space and model motion with a deformation field, it can reconstruct temporally consistent results, which is demonstrated in Fig. S5, where the single-image method ReEF [15] fails to produce consistent reconstruction on 2 input frames in a video sequence.

6. More Qualitative Results

6.1. More Results for Dynamic Reconstruction

Figure S6 demonstrates more dynamic reconstruction results on our method. Figure S7 shows the reconstruction results on a sequence with large motion. We can see that our method produces high-fidelity and temporally consistent dynamic garment reconstruction. More results can be seen in the attached video.

6.2. More Comparisons with Existing Methods

We show more visual comparisons between existing approaches and our method on PeopleSnapshot dataset [1] and our videos captured by smartphone in Fig. S8 and Fig. S9, respectively. Compared to state-of-the-art methods, our approach produces more realistic and delicate clothing surface, clearly demonstrating the effectiveness of our method.

More dynamic reconstruction comparison results can be found in our project page: https://lingtengqiu.github.io/2023/REC-MV/.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 3, 4

[2] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 1, 2

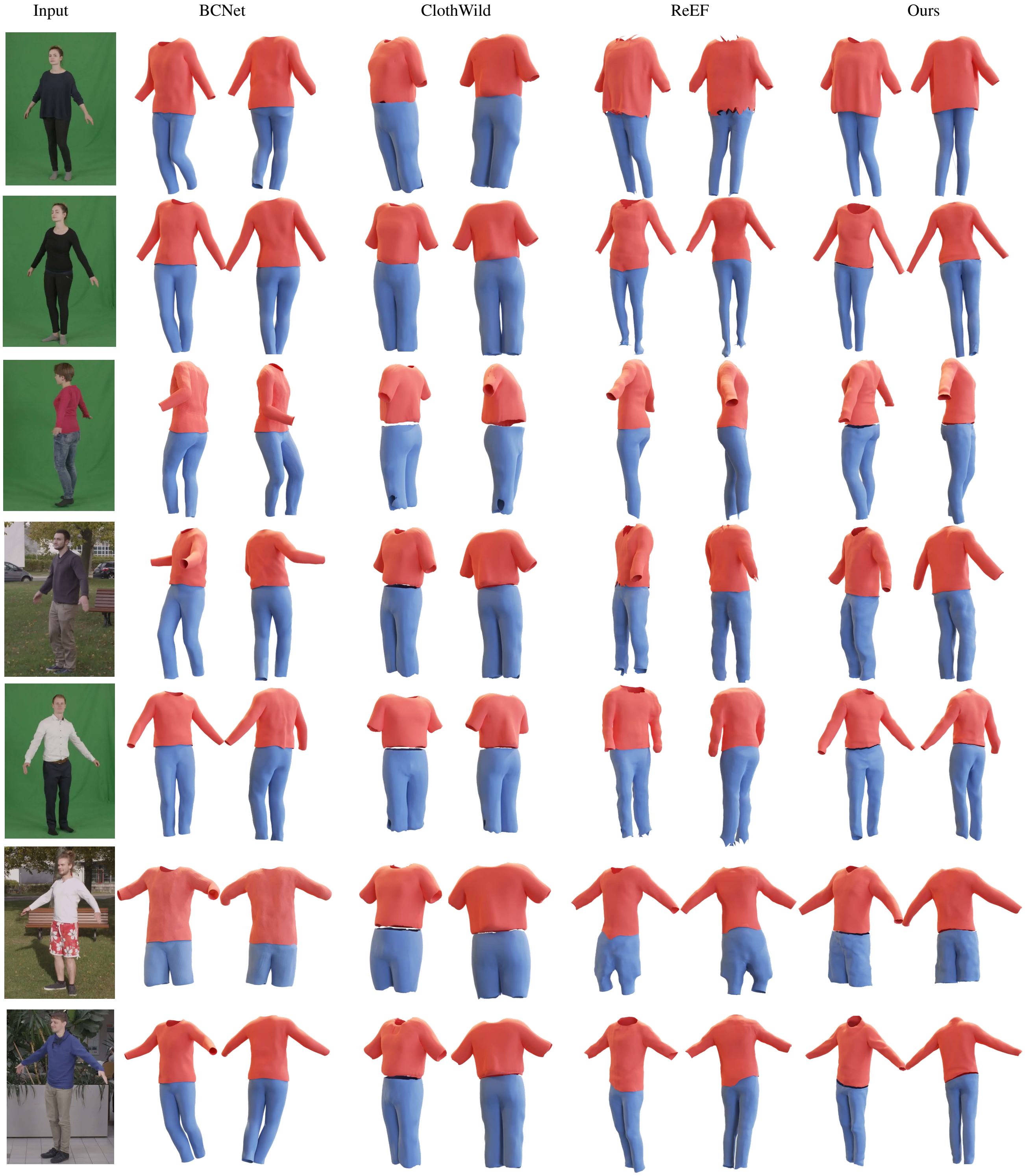

Figure S8. Reconstruction results on PeopleSnapshot. Each row shows the key frame of the video and results of different methods.

[3] Stuart Ganan and D McClure. Bayesian image analysis: An application to single photon emission tomography. *Amer. Statist. Assoc*, 1985. 2

[4] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *ICCV*, 2019. 1, 2

[5] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and

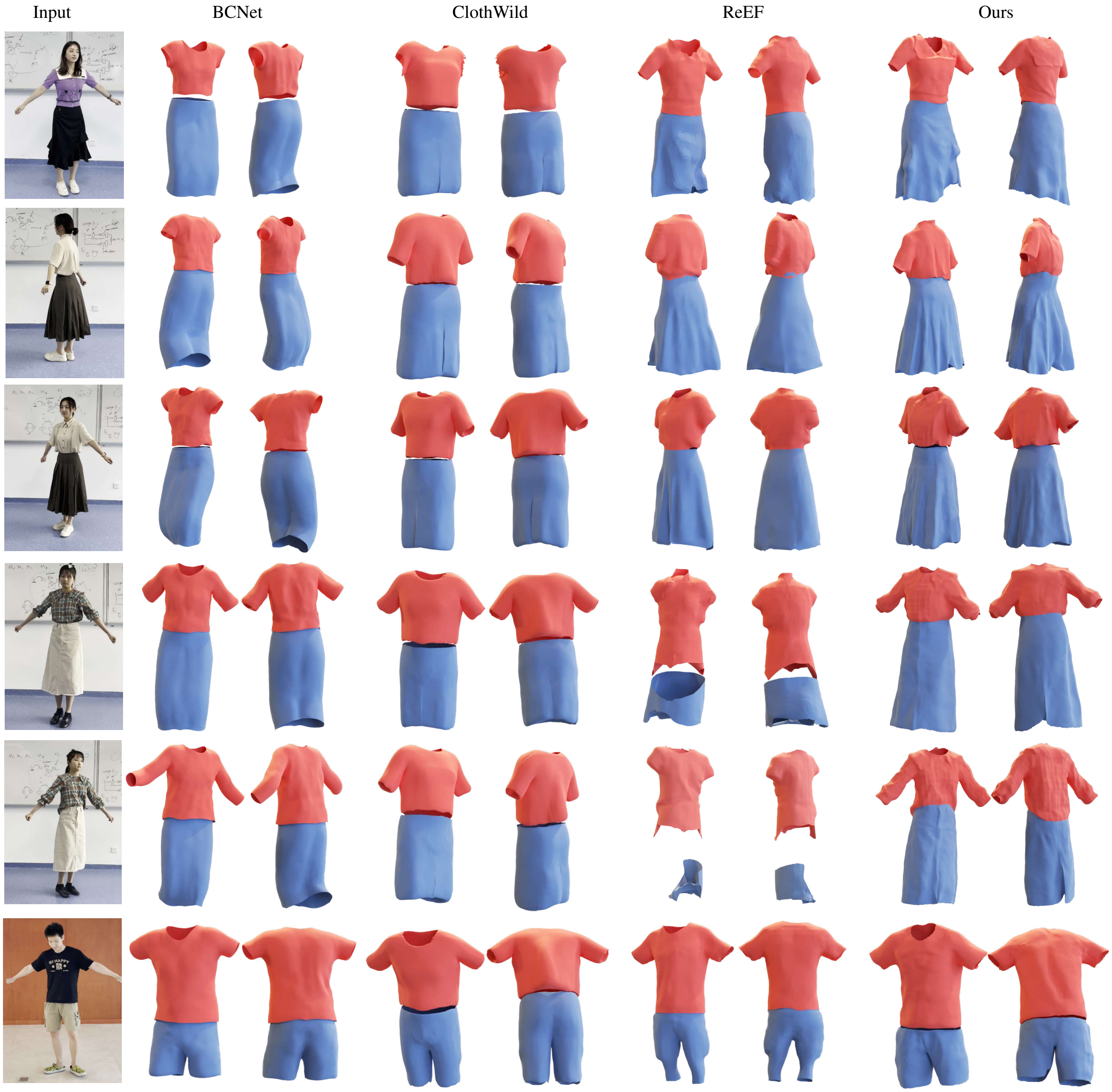

Figure S9. Results on videos taken by smartphones. Each row shows the key frame of the video and results of different methods.

Yaron Lipman. Implicit geometric regularization for learning shapes. In *Machine Learning and Systems*, 2020. 2

[6] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. 1, 2

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 2

[8] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *TPAMI*, 2020. 1

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[10] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH*, 1987. 2

[11] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 2

[12] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for

high-resolution 3d human digitization. In *CVPR*, 2020. 2

[13] Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. Laplacian surface editing. In *Eurographics/ACM SIGGRAPH symposium on Geometry processing*, 2004. 2

[14] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *ECCV*, 2020. 2

[15] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *CVPR*, 2022. 3, 4