Supplementary Material ReLoo: Reconstructing Humans Dressed in Loose Garments from Monocular Video in the Wild

Chen Guo*[1], Tianjian Jiang*[1], Manuel Kaufmann[1], Chengwei Zheng[1], Julien Valentin[2], Jie Song†[1], and Otmar Hilliges[1]

[1] ETH Zürich
[2] Microsoft

In this **supplementary document**, we provide additional materials to supplement our manuscript. In Sec. 1, we provide further implementation details of our proposed method ReLoo. Sec. 2 explains details of our experiment, including dataset descriptions and the implementation of baseline methods. Furthermore, in Sec. 3, we show additional quantitative and qualitative comparisons to demonstrate our superior performance over prior art and provide ablation studies of more components in our framework. Sec. 4 includes more qualitative results of our method on in-the-wild videos and an illustration of the decomposition of the inner body and outer garment layer. Finally, we discuss our limitations and potential negative societal impacts in Sec. 5. In the **supplementary video**, we show more 3D surface reconstruction and novel view synthesis results of dynamic humans dressed in loose outfits using our method and qualitative comparisons with baseline methods, both on indoor datasets and in-the-wild video sequences.

1 Implementation Details

1.1 Layered Neural Human Representation

Parameterization Details. In the main manuscript, we denote the body and garment layer with an implicit neural network f^B, or f^G respectively. In practice, f^B and f^G each consist of two separate neural networks that model the geometry and texture field of the respective layer, which is similar to [7,24]. For the neural body layer, our SDF network f_s^B that models the geometry takes the point $\boldsymbol{x}_c^B$ and the human pose parameters $\boldsymbol{\theta}$ as input and outputs the signed distance value s^B along with global geometry features $\boldsymbol{z}^B$ of dimension 256. Our body texture network f_c^B takes the point $\boldsymbol{x}_c^B$, the human pose parameters $\boldsymbol{\theta}$, points' normals $\boldsymbol{n}_d^B$ in deformed space, and the extracted 256-dimensional global geometry feature vectors $\boldsymbol{z}^B$ from the neural body SDF network as input and predicts the radiance value $\boldsymbol{c}^B$. Specifically, the points' normals $\boldsymbol{n}_d^B$ are calculated by the spatial gradient of the signed distance field f_s^B w.r.t. the 3D

[*] These authors contributed equally to this work
[†] Corresponding author. Now at HKUST(GZ) & HKUST

position in deformed space, following [7, 27]. This facilitates better disentanglement of surface geometry and appearance reconstruction. The same separation of geometry and texture field is applied to the neural garment layer f^G.

Network Architecture. The canonical neural body network f_s^B is modeled as an MLP with 8 fully connected layers, each of which consists of a weight normalization layer [22] and a Softplus activation layer. Each fully connected layer contains 256 neurons. Given the input point, we apply positional encoding with 6 frequency components to better model high-frequency details [17]. The canonical body texture network f_c^B is modeled as an MLP with 4 fully connected layers, each of which has the same architecture as the body shape network layers, except that it uses the Sigmoid activation function for the last layer and ReLU [19] for the rest of the layers. The network architectures for the neural garment model (f_s^G and f_c^G) follow the same approach.

1.2 Virtual-Bone Deformation

Parameterization Details. Instead of sticking to a pre-defined set of virtual bones with fixed positions and optimizing their transformations relative to the SMPL [15] root, we formulate the virtual bone deformation module using a deformation field $\mathcal{D}^G$. This allows for a progressively changing/updated garment topology during training and avoids overfitting to the learned garment shape after the first training stage, in which we only deploy skeletal deformation to drive both the inner body and outer garment layer (Sec. 3.4 in the manuscript). The virtual bone deformation field $\mathcal{D}^G$ takes the 3D positions $\boldsymbol{v}_i$ of the virtual bones along with conditions on the human body pose $\boldsymbol{\theta}$ and a continuous time embedding $\boldsymbol{t}$ as input. The continuous time embedding $\boldsymbol{t}$ is obtained by using the positional encoding introduced in NeRF [17] with 4 frequency components, which help to learn temporal dynamics from videos.

During the forward process of the virtual bone deformation, the skinning weights $\boldsymbol{\delta}_c$ (Sec. 3.2 in the manuscript) of $\boldsymbol{x}_c^G$ w.r.t. each virtual bone $\boldsymbol{v}_i \in \mathcal{V}$ is calculated based on the inverse of the distance between $\boldsymbol{x}_c^G$ and each $\boldsymbol{v}_i$. More specifically, we query the nearest K virtual bones from $\mathcal{V}$ for $\boldsymbol{x}_c^G$ based on the point-to-point distances in canonical space. The weights δ_c^i for the nearest K virtual bones are inversely proportional to the point-to-point distance and the skinning weights for the rest of the virtual bones are clamped to 0. We normalize $\boldsymbol{\delta}_c$ so that $\sum_{i=1}^{n_v} \delta_c^i = 1$. In our experiments, we set $K = 5$. To warp the sampled 3D garment points $\boldsymbol{x}_d^G$ in deformed space to canonical space, we first forward-warp the virtual bones' locations in canonical space $\boldsymbol{v}_i$ to deformed space using their own transformations $\boldsymbol{\mathcal{T}}_i$ and then proceed with the similar approach as the forward virtual bone deformation to obtain the skinning weights $\boldsymbol{\delta}_d$, and calculate the canonical correspondences $\boldsymbol{x}_c^G$.

Network Architecture. The virtual bone deformation field $\mathcal{D}^G$ for the garment layer is parameterized using an MLP with 4 fully connected layers, each of which consists of a weight normalization layer and a softplus activation layer. Each fully connected layer contains 256 neurons. We initialize the weights of the last layer of the deformation network to small values $\mathcal{U}(-10^{-5}, 10^{-5})$, *i.e.*, initializing the translations to be close to zero and the rotation matrices to be approximately identities.

1.3 Background Modeling and Scene Composition

Quadruple Reparameterization. We follow the inverted sphere parameterization of NeRF++ [26] to represent the background. Our human models are defined to be within a spherical inner volume with a radius equal to 3 and the background covers the complementary space. Specifically, each 3D background point $\boldsymbol{x}_d^S = (x_d^S, y_d^S, z_d^S)$ is reparametrized by the quadruple $\boldsymbol{x'}_d^S = (x'^S_d, y'^S_d, z'^S_d, \frac{1}{r})$, where $\left\| \left(x'^S_d, y'^S_d, z'^S_d \right) \right\| = 1$ and $(x_d^S, y_d^S, z_d^S) = r \cdot (x'^S_d, y'^S_d, z'^S_d)$. Here r denotes the magnitude of the vector from the camera origin to $\boldsymbol{x}_d^S$. This reparameterization of the background points helps to improve the numerical stability and to weigh further away points with lower resolution. To obtain the background component RGB value, we follow NeRF++ and sample 32 background points. This is done by uniformly sampling $\frac{1}{r}$ in the range $[0, \frac{1}{3}]$, where 3 corresponds to the predefined inner volume radius. Given the sampled $\frac{1}{r}$, we calculate the corresponding background point $\boldsymbol{x'}_d^S$ using the geometric relationship derived in [26].

Scene Composition. To obtain the final rendered pixel value, we raycast the human layers and the background volume separately and composite the rendered color of humans $\hat{C}^H$ with the one of the scene background $\hat{C}^S$. The final pixel color value is calculated by:

$$C = C^H + (1 - \hat{O}^H)\ C^S, \tag{16}$$

where $\hat{O}^H = \sum_{i=1}^{N} \sum_{p=1}^{P} \left[o_i^p \prod_{q=1}^{P} \prod_{j \in \mathcal{Z}_i^{q,p}} \left(1 - o_j^q\right) \right]$ is the total opacity for all the person in the scene, and we follow the same notations as our manuscript.

Network Architecture. The scene background network f^S consists of two parts: the density network and the texture network. The density network has the same architecture as the canonical human shape network with 10 frequency components to the input background points. The texture network only includes 1 block of a fully connected layer with 128 neurons, a weight normalization layer, a ReLU activation layer, and a Sigmoid activation layer at the end. Both the density network and the texture network take the quadruple parameterization of the sampled background point, view direction, and per-frame learnable time encoding as input and output the density and the view-dependent radiance value. The per-frame time encoding helps to compensate for dynamic changes in the environment.

1.4 Preprocessing

Pose Initialization. To obtain pose initialization for in-the-wild videos, we first leverage 4DHumans [5] to estimate the SMPL [15] parameters. However, [5] usually assumes an extremely large focal length which would make the 3D human extremely small within the sphere (radius equal to 3) after the camera normalization. Specifically, the same as HMAR [21], in 4DHuman, the SMPL translation $\boldsymbol{T}_{\text{SMPL}}$ in the camera frame can be re-calculated using the following formula:

$$\boldsymbol{T}_{\text{SMPL}} = \left[t_x + \frac{2c_x - 2p_x}{sb}, t_y + \frac{2c_y - 2p_y}{sb}, \frac{2f}{sb} \right] \tag{17}$$

where t_x, t_y, and s represent the predicted local camera parameters related to the bounding boxes, b. c_x and c_y correspond to the scale and center of the respective bounding box of the target. (p_x, p_y) and f represent the principal point (x, y) and focal length of the camera, respectively. Throughout our experiment, we use the following values for the principal point and focal length of the camera

$$p_x = \frac{W}{2}, p_y = \frac{H}{2}, f = \frac{W + H}{2} \tag{18}$$

where W and H denote the image width and height. After this conversion, we obtain a set of aligned camera and SMPL parameters with a reasonable camera focal length, which empirically helps to provide better results.

Segmentation Mask Initialization. We build a semi-automatic preprocessing pipeline to estimate full-body human and clothing segmentation masks using SAM [10,13]. We first utilize Graphonomy [6] to parse the human in the images, yielding coarse segmentation masks. Given the outputs from Graphonomy, we concatenate sampled pixels from the body or clothing classes with corresponding 2D keypoints detected using OpenPose [1]. The concatenation serves as the input point and mask prompts for SAM to predict finer segmentation masks for the body and garment. However, the segmentation masks do not consistently exhibit robustness, and the delineations of boundaries might lack sharpness. Manual annotations such as manually selecting pixels to serve as extra point prompts are occasionally required to further improve the segmentation mask quality.

1.5 Multi-Garment Clothed Human

Our method also generalizes to multiple garments, *e.g.*, a T-shirt for the upper body and a skirt for the lower body (*cf*. Fig. 3 in manuscript and Fig. 12). We separately model these two garments with two neural networks to keep the representation capacity. Moreover, considering that T-shirts and other upper-body garments mainly follow skeletal movement, we thus use SMPL-based skeletal deformation to warp sampled 3D points for the upper-body garment layer, while the lower-body garment is still deformed using the proposed virtual bone deformation module.

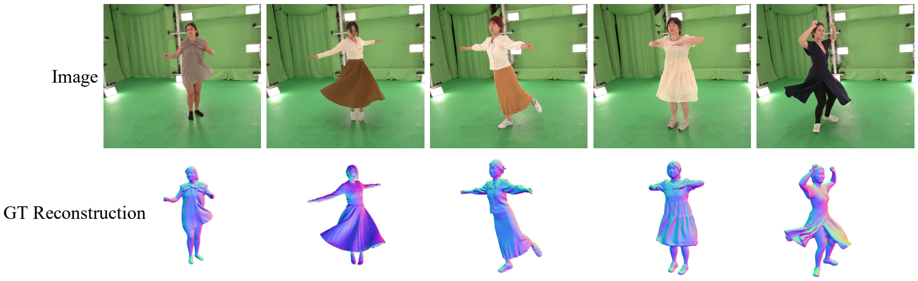

Fig. 9: MonoLoose Dataset. We show sample images and their corresponding ground-truth meshes from MonoLoose dataset.

1.6 Inner Body Regularization Loss

Due to garment-body occlusions, reconstructing a plausible inner human body is an ill-posed problem due to the lack of observations. To stabilize the training progress while preserving a reasonable inner body shape, in addition to the training objectives that are described in the manuscript, we include a body loss $\mathcal{L}_{\text{body}}$ that encourages the neural inner body to be close to the optimized SMPL shape. We gradually decay the weight of this loss as it is especially needed in the early stage of the training.

1.7 Training Details

We initialize both the canonical body shape and the garment shape networks with a generic SMPL body by using a subset of motion sequences released in AMASS [16]. Human-specific shape network initialization can accelerate the training convergence. We optimize our neural networks and pose parameters using the Adam optimizer [12]. The learning rate for training our neural networks is set to $l = 5e^{-4}$ and the learning rate for optimizing the pose parameters is set at one-tenth of l initially. We decay the learning rates in half after 200/500/800/1000 epochs respectively. The other Adam hyper-parameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We train an individual model for every input video. A video with about 300 frames (10 seconds) usually requires training for 2 days on a single NVIDIA 3090Ti (24GB) with batch size 512.

2 Evaluation Details

2.1 MonoLoose Dataset

We collect a new dataset called MonoLoose, which has a particular focus on human subjects dressed in loose attire while performing highly dynamic motions. MonoLoose is captured with a high-end dense-view camera rig. The dataset contains 5 sequences with different identities, loose garment styles, and motions (in total 1219 frames). This dataset is specifically curated for evaluating monocular human surface reconstruction and novel view synthesis methods with the provided high-fidelity 3D ground-truth meshes. These dense human meshes are reconstructed by 106 synchronized cameras (53 RGB and 53 IR cameras) via commercial software [2]. We use two separate cameras for training and evaluation of novel view synthesis (The distances between the test and training view are 0.56 meters and 18 degrees). We show example images and the corresponding 3D ground-truth meshes of our MonoLoose dataset in Fig. 9.

Ethics. Our institution's ethics committee duly approved the protocol we followed for the collection and publication of MonoLoose dataset. All subjects have freely volunteered to participate in this data collection. They have been duly informed about the intended use and publication of the dataset, signed a consent form, and have received compensation for the time it took to record them.

2.2 DynaCap Dataset

For the novel view synthesis evaluation on DynaCap [8], we curate 2 sequence clips (in total 620 frames), *i.e.*, FranziBlue and FranziRed, where we regard images from camera 17 as training camera view and camera 16 as test view. The camera distances for the DynaCap experiments are 0.7 meters and 16 degrees.

2.3 Baseline Methods

In our manuscript, we compare our method with state-of-the-art video-based methods (SelfRecon [9], Vid2Avatar [7], and SCARF [3]) in two tasks: 3D human surface reconstruction and novel view synthesis. SelfRecon and Vid2Avatar are SMPL-based methods, the same as ReLoo, and share the same SMPL model parameters for training and testing. SCARF takes SMPL-X parameters as input, thus, we first utilize the officially released model transfer scripts [20] to convert SMPL to SMPL-X models with the corresponding parameters. We use the converted model parameters to serve as the input for SCARF for training and testing.

3 Additional Experimental Results

3.1 Surface Reconstruction Comparisons

We provide additional qualitative reconstruction comparisons with SelfRecon [9], Vid2Avatar [7] and SCARF [3] in Fig. 10. Compared to state-of-the-art video-based human reconstruction methods in both categories (single-layer and multi-layer), our method outperforms them by a large margin both quantitatively (*cf*. Tab. 1 in manuscript) and qualitatively (*cf*. Fig. 10). Our approach can accurately reconstruct complete and more detailed 3D human surfaces, and capture large non-rigid garment surface deformations.

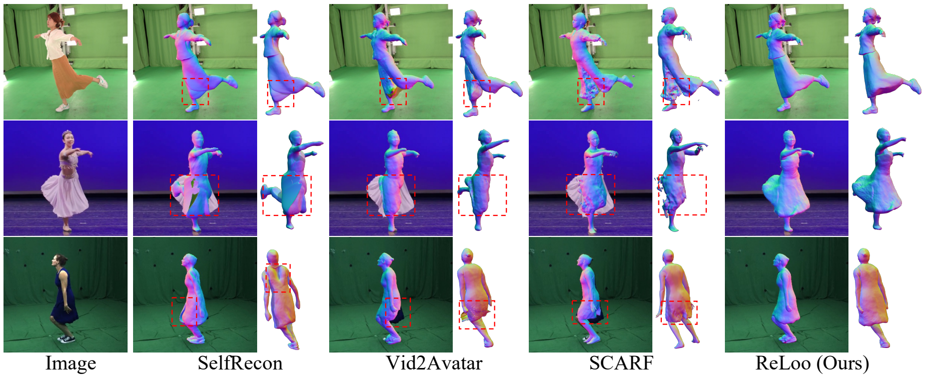

Fig. 10: Qualitative 3D surface reconstruction comparison. Baseline methods produce less detailed and implausible 3D clothed human reconstructions with visible artifacts (discontinuities between legs, missing dress parts) due to the strong reliance on skeletal deformations. In contrast, our method correctly recovers the clothing dynamics and generates more detailed and complete 3D human surfaces. Note also that ReLoo produces more detailed facial features.

3.2 Novel View Synthesis Comparisons

We show additional qualitative novel view synthesis comparisons in Fig. 11. Our method outperforms baseline methods both quantitatively (*cf*. Tab. 2 in the manuscript) and qualitatively (*cf*. Fig. 11). ReLoo produces more plausible and realistic renderings while preserving sharper and fine-grained texture details.

3.3 Surface Reconstruction Comparisons with Image-based Method

In our manuscript, we mainly compare with video-based 3D human reconstruction for a fair comparison. Here, we complement our surface reconstruction comparison experiments with an image-based baseline method ECON [23]. ECON is a state-of-the-art regression-based model for reconstructing 3D humans from images, capable of handling humans dressed in loose garments. As indicated in Tab. 3, our method outperforms ECON by a substantial margin on the MonoLoose dataset. Especially, our Chamfer distance error is only about 40% of ECON's error. This performance difference is even more visible in qualitative comparisons shown in Fig. 12. When loose garments exhibit large non-rigid surface deformations during human articulation, ECON fails to recover complete 3D human shapes but only outputs corrupted reconstructions (*e.g.*, missing body parts and clothing). Furthermore, ECON fails to preserve fine-grained surface details on the reconstructions (*e.g.*, the human face in the second row of Fig. 12). In contrast, our method generates complete and high-fidelity 3D human surfaces even when loose outfits show extremely free-form deformations.

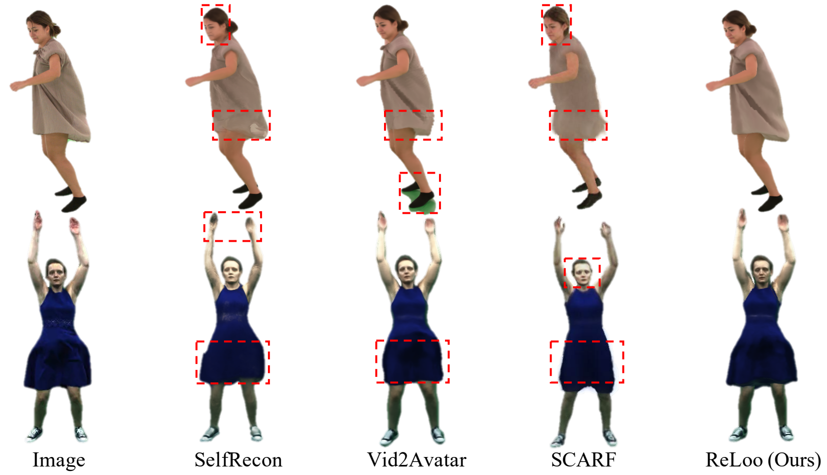

Fig. 11: Qualitative novel view synthesis comparison. Our method achieves better rendering quality with clearly sharp boundaries and detailed texture recovery in *e.g.*, garment patterns and faces. Baseline methods can only produce corrupted and blurry rendering results (dress discontinuities between legs and unsharp texture details).

Table 3: Quantitative evaluation on surface reconstruction with image-based method. We compute the 3D surface metrics on the MonoLoose dataset. Our method consistently outperforms ECON on all evaluation metrics (*cf*. Fig. 12).

Method	$\mathbf{C}-\ell_2\downarrow$	$\mathbf{NC}\uparrow$	$\mathbf{V-IoU}\uparrow$
ECON [23]	4.49	0.688	0.695
Ours	**1.93**	**0.831**	**0.881**

3.4 Qualitative Comparisons with GS-based Method

We further provide some qualitative comparisons with concurrent Gaussian-Splatting (GS) [11] based human reconstruction methods on the UBC-Fashion dataset [25]. GART [14] is one of the concurrent GS-based works and it leverages the explicit and efficient representation to achieve human reconstruction with fast training and real-time inference speed. We show the learned canonical human models of GART and our method ReLoo in Fig. 13. Compared to our implicit SDF-based representation, the explicit Gaussian-based representation does not demonstrate plausible 3D human shapes due to the irregularity of the 3D Gaussians' distribution. GS-based methods can easily fit the image observations without preserving a plausible 3D shape, leading to visually unpleasant rendering results with scaly artifacts.

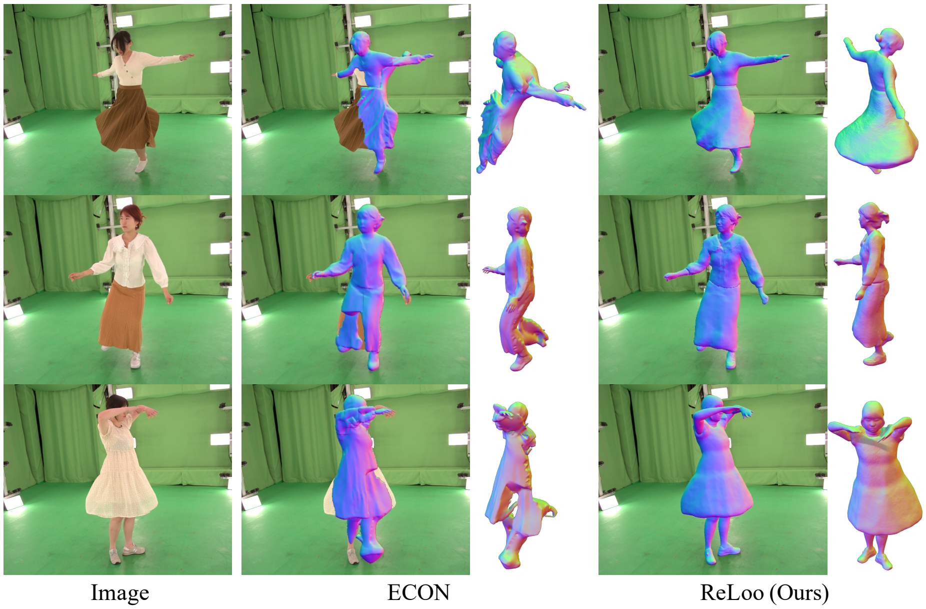

Fig. 12: Qualitative comparisons with image-based method. Compared to the state-of-the-art image-based method ECON [23], our representation and learning schemes enable more robust and detailed human surface reconstructions when dressed in highly dynamic loose garments.

3.5 Robust Error Function *vs.* L_1 Loss

The segmentation masks obtained from SAM [10, 13] are not perfect and the predictions might be noisy. Directly applying L_1 loss as the segmentation mask objective might lead to unstable training progress since the segmentation mask predictions can be contradictory across video frames. Instead of using a hard L_1 loss for supervision, we employ a robust Geman-McClure error function ρ [4] which helps to down-weigh potentially erroneous cloth segmentation mask predictions. Moreover, the inverse CDF sampling process [24] does not always guarantee a spiky distribution of the ray-sampled points around surface boundaries. When sampled points for the neural body layer are interwoven with the sampled points for the neural garment layer, the ray opacities of the garment layer $\hat{O}^G(\boldsymbol{r})$ cannot be strictly equal to 1. A hard constraint introduced by L_1 loss can lead to corrupted surface reconstructions (thicker garment layer). We provide a qualitative comparison in Fig. 14 to highlight the effects of the robust error function compared to a simple L_1 loss. Our method recovers more realistic and detailed 3D surface reconstructions with the robust Geman-McClure error function.

3.6 Layered Neural Human Representation

Our layered neural human representation is not only the foundation for reconstructing humans dressed in loose garments but can also contribute to recon-

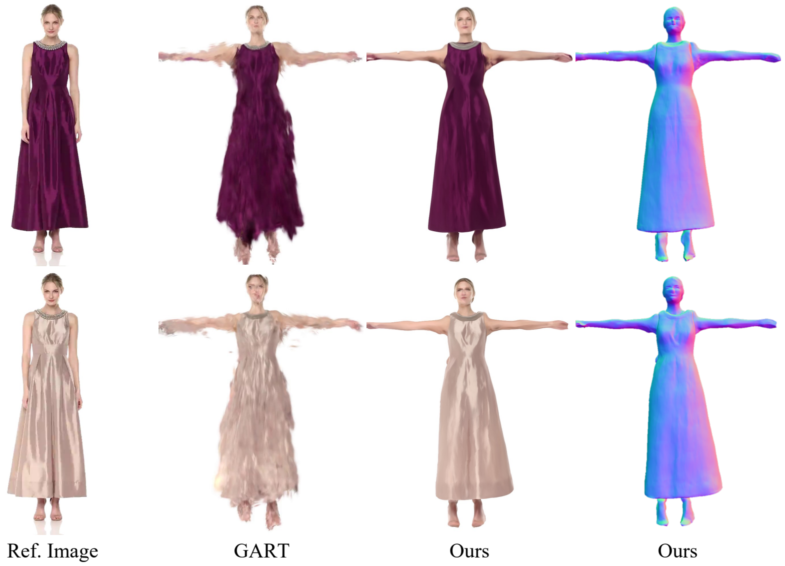

Fig. 13: Qualitative comparisons with GS-based method. Compared to the GS-based method, our implicit SDF-based representation can better learn a plausible 3D human shape with detailed geometry and appearance.

structing more detailed humans. To validate this, we compare our multi-layer design to a single-layer representation on the SynWild dataset [7] which consists of human subjects with tight-fitting clothing. Note that in this experiment, we deactivate the virtual bone deformation module as tight-fitting garments mostly follow skeletal movement. As illustrated in Fig. 15, our layered human representation recovers finer human surface details, such as faces and clothing wrinkles, which the single-layer representation struggles with, even when both models have the same number of parameters. This can be naturally explained by the observation that implicit neural networks can better learn more local features with a decomposed representation, *i.e.*, the neural body layer can focus more on body shape features without having to spend capacity on garment details.

4 Visualization

As shown in Fig. 16, our method ReLoo can generalize to humans with different body shapes, miscellaneous loose garment styles, and diverse human motions. ReLoo can even reconstruct extremely challenging outfits such as a very wide cloak in full swing (*cf*. the first column of Fig. 16). We also demonstrate the decomposition of the learned body and garment layer separately in Fig. 17. ReLoo achieves an accurate decomposition of the inner body and outer clothing, as well as high-fidelity 3D reconstructions for both layers from only monocular inputs.

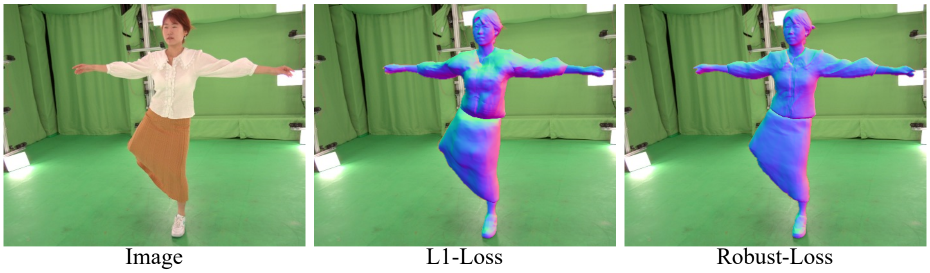

Fig. 14: Importance of robust loss. Our method recovers more realistic and detailed 3D surface reconstructions with the robust Geman-McClure error function for segmentation mask supervision.

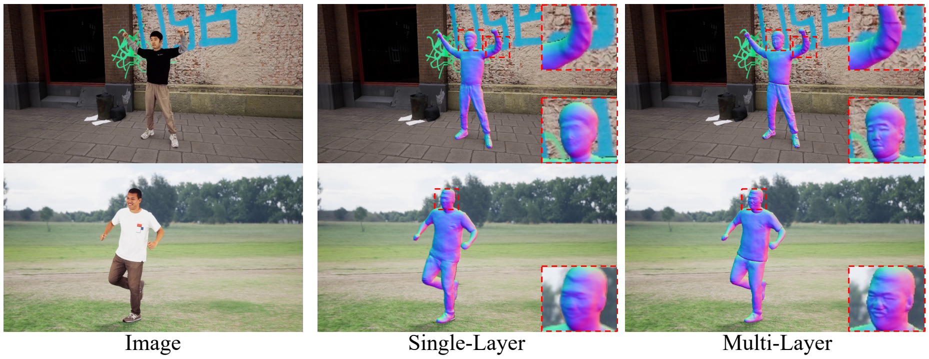

Fig. 15: Importance of layered neural human representation. Our layered human representation can recover more fine-grained human surface details (*e.g.*, faces and clothing wrinkles).

5 Limitations and Societal Impact Discussion

Although readily available, our method ReLoo still relies on reasonable pose estimates and segmentation masks as inputs. As mentioned in Sec. 1.4, the segmentation process is a semi-automatic pipeline and manual efforts for annotation are occasionally required to improve the quality of the segmentation mask. Currently, our method is mainly deployed to up to two garments. The computational complexity of ReLoo increases linearly with the number of garments that we aim to reconstruct separately, making it less scalable to cases where we aim to reconstruct the clothed human with various accessories (such as hats, gloves, and *etc.*) in layers. Future work could incorporate recent advances in fast and memory-efficient representation [11, 18] to achieve highly efficient layered representation. Furthermore, our method doesn't explicitly model hands, leading to less detailed hand reconstructions (*cf.*, Fig. 12). We believe the integration of an expressive human model such as SMPL-X [20] is a promising future direction.

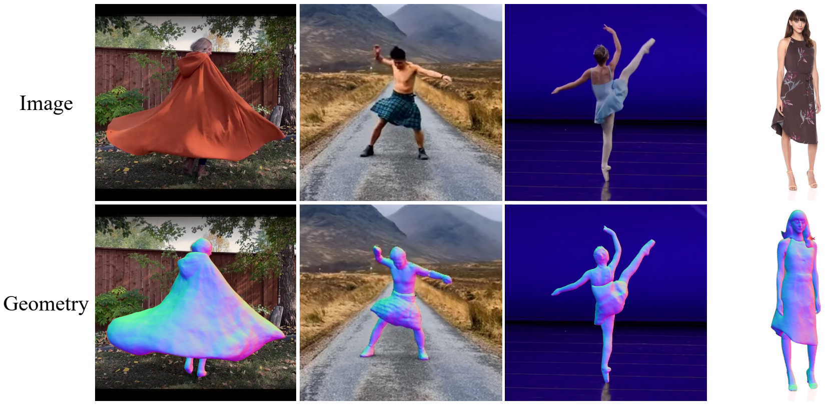

Fig. 16: Additional qualitative results. Our method ReLoo generalizes to humans with different body shapes, miscellaneous loose garment styles, and diverse human motions.

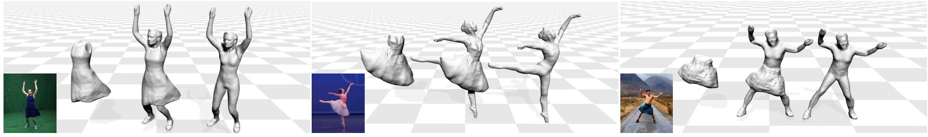

Fig. 17: Decomposition of body and garment. Our method accurately decomposes the clothed human into the inner body and outer clothing, achieving high-fidelity 3D reconstruction results.

ReLoo enables high-fidelity digitization of humans from a single monocular in-the-wild video, which bears the potential to facilitate diverse downstream applications in the film and gaming industries, as well as AR/VR environments. The final outcome of ReLoo is realistic digital avatars, capable of being animated to novel poses given respective input signals. This may lead to concerns regarding privacy leaks and the potential for the misuse of digital assets, for example by creating digital avatars from people who did not consent to such uses and subsequent misappropriation of these avatars for dubious purposes. When developing avatar creation methods, be it for research or products, paramount focus should be directed towards addressing these concerns. We strive towards enabling utilization of such technology in manners that are advantageous for society. Regrettably, the prevention of malevolent applications can however not be fully guaranteed. We posit that prioritizing a comprehensive and transparent treatment of these methodologies – including discussion of technical details within the paper along with the provision of code and data – should take precedence over undisclosed research. This approach is essential for devising effective countermeasures aimed at mitigating the potential for unethical applications.

References

1. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Real-time multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
2. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Trans. Graph. **34**(4) (jul 2015). `https://doi.org/10.1145/2766945`, `https://doi.org/10.1145/2766945`
3. Feng, Y., Yang, J., Pollefeys, M., Black, M.J., Bolkart, T.: Capturing and animation of body and clothing from monocular video. In: SIGGRAPH Asia 2022 Conference Papers. SA '22 (2022)
4. Geman, S., McClure, D.E.: Statistical methods for tomographic image reconstruction (1987)
5. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa*, A., Malik*, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: International Conference on Computer Vision (ICCV) (2023)
6. Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., Lin, L.: Graphonomy: Universal human parsing via graph transfer learning. In: CVPR (2019)
7. Guo, C., Jiang, T., Chen, X., Song, J., Hilliges, O.: Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
8. Habermann, M., Liu, L., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Real-time deep dynamic characters. ACM Transactions on Graphics **40**(4) (aug 2021)
9. Jiang, B., Hong, Y., Bao, H., Zhang, J.: Selfrecon: Self reconstruction your digital avatar from monocular video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
10. Ke, L., Ye, M., Danelljan, M., Liu, Y., Tai, Y.W., Tang, C.K., Yu, F.: Segment anything in high quality. In: NeurIPS (2023)
11. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2015)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
14. Lei, J., Wang, Y., Pavlakos, G., Liu, L., Daniilidis, K.: Gart: Gaussian articulated template models (2023)
15. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015)
16. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)
17. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)

18. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022). https://doi.org/10.1145/3528223.3530127, https://doi.org/10.1145/3528223.3530127
19. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. p. 807–814. ICML'10, Omnipress, Madison, WI, USA (2010)
20. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
21. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people with 3d representations. In: NeurIPS (2021)
22. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016), https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf
23. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: ECON: Explicit Clothed humans Optimized via Normal integration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
24. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Advances in Neural Information Processing Systems (2021)
25. Zablotskaia, P., Siarohin, A., Zhao, B., Sigal, L.: Dwnet: Dense warp-based network for pose-guided human video generation. In: 30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019. p. 51. BMVA Press (2019)
26. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv:2010.07492 (2020)
27. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: I M Avatar: Implicit morphable head avatars from videos. In: Computer Vision and Pattern Recognition (CVPR) (2022)