




# Learning 3D Human UV with Loose Clothing from Monocular Video

Meng-Yu Jennifer Kuo<sup>1</sup><sup>a</sup>, Jingfan Guo<sup>1</sup><sup>b</sup> and Ryo Kawahara<sup>2</sup><sup>c</sup>

<sup>1</sup>University of Minnesota, U.S.A.

<sup>2</sup>Kyushu Institute of Technology, Japan

**Keywords:** 3D Human Reconstruction, UV Mapping, Single View.

**Abstract:** We introduce a novel method for recovering a consistent and dense 3D geometry and appearance of a dressed person from a monocular video. Existing methods mainly focus on tight clothing and recover human geometry as a single representation. Our key idea is to regress the holistic 3D shape and appearance as a canonical displacement and albedo maps in the UV space, while fitting the visual observations across frames. Specifically, we represent the naked body shape by a UV-space SMPL model, and represent the other geometric details, including the clothing, as a shape displacement UV map. We obtain the temporally coherent overall shape by leveraging a differential mask loss and a pose regularization. The surface details in UV space are jointly learned in the course of non-rigid deformation with the differentiable neural rendering. Meanwhile, the skinning deformation in the garment region is updated periodically to adjust its residual non-rigid motion in each frame. We additionally enforce the temporal consistency of surface details by utilizing the optical flow. Experimental results on monocular videos demonstrate the effectiveness of the method. Our UV representation allows for simple and accurate dense 3D correspondence tracking of a person wearing loose clothing. We believe our work would benefit applications including VR/AR content creation.

## 1 INTRODUCTION

3D human shape reconstruction is crucial as it finds applications in a wide range of domains including 3D avatars in games and metaverse, as well as virtual fitting. Various approaches have been proposed for this study. Specifically, there are methods using videos captured by a large number of perfectly calibrated cameras (Zhao et al., 2022; Wang et al., 2022), and methods that recover the 3D shape by refining the captured depth (Newcombe et al., 2015). Most of the images captured by surveillance cameras and on the Internet, however, are monocular images. Methods that require specialized capture environments limit the utility at the consumer level. Recently, several methods have been introduced to recover the 3D shape of a person from a monocular video by optimizing a parametric human model (Guo et al., 2023), and have achieved compelling results.


Although parametric human model such as SCAPE (Anguelov et al., 2005) and SMPL (Loper et al., 2015) leans a powerful means for accurate 3D human modeling, these methods are mainly limited in two critical ways. First, they mainly focus on the





Figure 1: Our method achieves holistic, temporally coherent 3D dressed human reconstruction from a monocular video. Our method also realizes dense surface correspondence tracking over the sequence.

human wearing tight clothing. This assumption hinders the utility especially for a person wearing skirts or dresses. Most importantly, most of these methods are limited to recovering the geometry as a single representation. This could be a deal-breaker for some applications, including virtual try-on, where having 3D human models in which the garment can be modified with different textures and/or shapes is crucial.

In this work, we propose a novel method to create the 3D avatar of a person wearing loose clothing from a monocular video. Our key idea is to regress the holistic 3D shape and appearance as canonical UV-

<sup>a</sup> <https://orcid.org/0000-0002-6705-7971>

<sup>b</sup> <https://orcid.org/0009-0008-6198-365X>

<sup>c</sup> <https://orcid.org/0000-0002-9819-3634>

space shape displacement and albedo maps while fitting the visual observations across frames. We represent the naked body shape by a standard-resolution SMPL model (Loper et al., 2015) in the UV space using UV mapping (Blinn and Newell, 1976), and assume the model detail (including clothing and hair) is a sub-map of the canonical UV map. Such UV representation provides a mapping between each 3D vertex and a predefined 2D space. The shape displacement UV map encodes the freeform offsets. We use these UV maps to augment the naked SMPL.

We utilize differential mask loss and a pose regularization to obtain the temporally coherent overall shape. The details on the surface in UV space are jointly refined with the differentiable neural rendering. To achieve better rendering, we decompose RGB images to obtain the diffuse albedo, and further refine light source and camera view directions. Meanwhile, the skinning deformation in the garment region is updated periodically to adjust its residual non-rigid motion in each frame. We also leverage optical flow to obtain temporally consistent representations over the sequence, and a symmetric structure constraint is enforced to better account for the invisibility.

We quantitatively and qualitatively evaluate our method on both synthetic and real video datasets, as well as on Internet videos, with a subject wearing loose clothing. We regress the canonical UV representation for each subject in a self-supervision manner. Experimental results effectively demonstrate that our pixel-aligned UV prediction achieves full (fuller) and dense reconstruction of the target person. We also show that our method realizes dense surface correspondence tracking over the sequence, enabling re-texturing and/or garment transfer. We believe that our work would expand the application of 3D human generation in a wide range of fields.

## 2 RELATED WORKS

**Holistic Human Reconstruction from Multi-View/Depth.** In general, 3D reconstruction requires multi-view image data, to enable triangulation. The number of cameras required to reconstruct fine-grained geometries is usually very high (Joo et al., 2015). There are several approaches using multi-view RGB (Zhao et al., 2022; Wang et al., 2022; Hilton and Starck, 2004) or RGBD (Dong et al., 2022) cameras to capture full human body. In real-world scenarios, however, sometimes it is difficult to install that many cameras, perhaps 2 or 3 at most, or perhaps only one camera. Requiring a multi-view capture system greatly limits the application of these methods.

For depth-based approaches, a pioneering work by Newcombe *et al.* (Newcombe et al., 2015) proposed depth refinement through integration of 3D volumes across time. While the aforementioned approaches have yielded compelling results, they still require specialized setup of the capture system and are therefore not user-friendly at the consumer level.

**Holistic Human Reconstruction from Monocular Video.** For single-view human reconstruction (Alldieck et al., 2019), and synthetic data generation (Varol et al., 2017), parametric 3D human models such as SCAPE (Anguelov et al., 2005) and SMPL (Loper et al., 2015) are widely used. Extending such parametric models to generate 3D clothing or clothed humans could be challenging (Ma et al., 2020). For single-image approaches, Tex2Shape (Alldieck et al., 2019) represented geometry as displacements in UV space to the surface of the SMPL body model. However, it only estimates the shape of observed subject and is limited to tight clothing. In our work, we also adopt similar UV representation but go beyond it in terms of reconstructed surface properties (albedo) and in terms of reconstructed clothing (dresses and skirts).

Recent works on regressing 3D surfaces from images have shown promising results (Xiu et al., 2023; Alldieck et al., 2022). These methods, however, require high-fidelity 3D data for supervision, and they only recover the geometry at one time instance thus cannot represent a temporally coherent shape reconstruction over the entire sequence. Recently, several methods proposed to obtain articulated human models by fitting implicit neural fields to video via neural rendering while requiring external segmentation methods (Jiang et al., 2022; Weng et al., 2022). Vid2Avatar (Guo et al., 2023), on the other hand, jointly solves scene decomposition and 3D reconstruction. While these methods achieve compelling results, they are fundamentally limited to tight clothing and/or single geometry representation.

## 3 METHOD

Given a monocular video of a person, our goal is to learn its full-body model with realistic appearance and geometry in the UV space, while enabling garment transfer and re-texturing. An overview of our method is shown in Fig. 2 and Fig. 4.

### 3.1 Canonical Human Generator

We parameterize canonical human in the UV space by leveraging a human shape prior in the form of T-posed

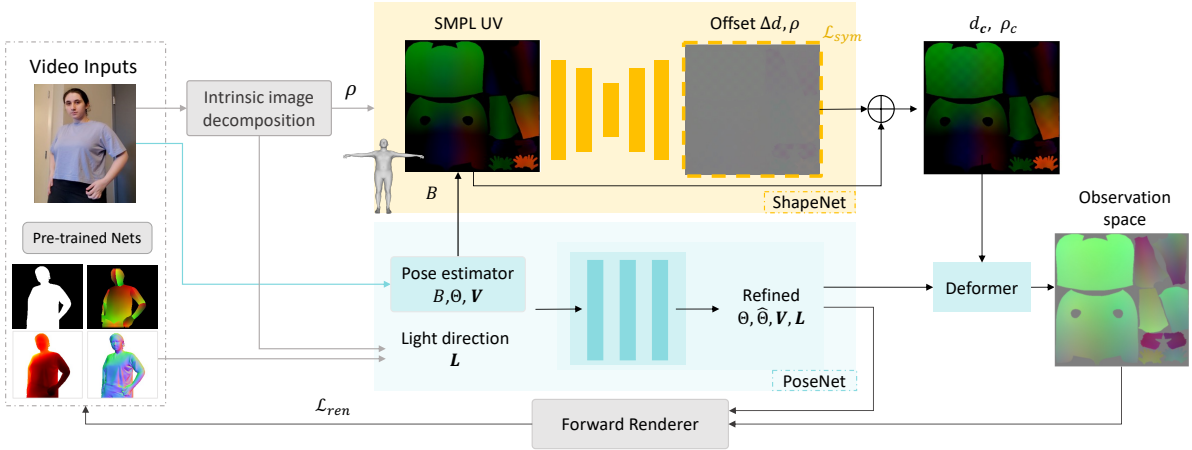


Figure 2: Method overview. Given a monocular video of a person, our method optimizes for the canonical albedo  $\rho_c$  and geometry  $d_c$  in the UV space, light source directions  $L$ , camera viewing directions  $V$  as well as the motion field:  $\{\Theta, \hat{\Theta}\}$  transforming from the canonical to the observation space.

naked SMPL (Loper et al., 2015) using UV mapping (Blinn and Newell, 1976). For each query point  $\mathbf{x}$  in the canonical UV space, we predict a shape displacement vector  $\Delta d(\mathbf{x}) \in \mathbb{R}^3$  from the base model  $d_{base}(\mathbf{x})$  to model details including the clothing, as well as its diffuse albedo  $\rho(\mathbf{x}) \in \mathbb{R}^3$  as follows

$$\begin{aligned} d_c(\mathbf{x}) &= d_{base}(\mathbf{x}) + \Delta d(\mathbf{x}), \\ \rho_c(\mathbf{x}) &= \rho(\mathbf{x}). \end{aligned} \quad (1)$$

We augment naked SMPL with geometric and appearance details using these two UV maps, including shape and texture UV maps:  $\{\Delta d(\mathbf{x}), \rho(\mathbf{x})\}$  in canonical space  $\mathbf{c} : \mathbb{R}^{H \times W} \times \mathbb{R}^6$ , where  $H \times W$  is the resolution of UV map. Note that the details of mesh model is proportional to the resolution of the UV map (Alldieck et al., 2019).

### 3.2 Deformer

In order to learn the canonical UV model map from posed images, we need the appearance and the 3D geometry in the observation space.

**Shape Deformation.** Given bone pose parameters  $\theta \in \mathbb{R}^{3 \times 24}$ , we transform each canonical point  $\mathbf{x}$  into the observation space using Linear Blend Skinning  $T(\cdot)$ :

$$\hat{\mathbf{x}} = T(\mathbf{x}, \theta, w) = \sum_{i=1}^K w_i(\mathbf{x}) \mathcal{B}_i(\mathbf{x}, \theta_i), \quad (2)$$

where  $\mathcal{B}_i$  and  $w_i$  are the transformation and the canonical blend weight for  $i$ -th bone, respectively, and  $K$  is the number of joints.

The weights are nonzero and affect each canonical point  $\mathbf{x}$ . To avoid redundant blend weights, we

represent canonical blend weight by interpolating the weight  $\hat{w}_i$  assigned to each vertex of the mesh as

$$w_i(\mathbf{x}) = \sum_{j=1}^3 \lambda_j(\mathbf{x}) \hat{w}_i(\mathbf{m}_j), \quad (3)$$

where  $\mathbf{m}_j$  denotes the  $j$ -th vertex of the face to which the point  $\mathbf{x}$  belongs, and  $\lambda_j$  denotes the interpolation weight of the  $j$ -th vertex in the barycentric coordinate system (Floater, 2003).

**Garment Deformation.** We assume the model detail (clothing, hair, and shoes) is a sub-map of the canonical UV map, and assume each sub-map point  $\mathbf{x}_g$  is associated with a body point  $\mathbf{x}$ . That is, the deformation in the garment region is conditioned to the shape deformation. We articulate each sub-map point  $\mathbf{x}_g$  as:

$$\hat{\mathbf{x}}_g = T(T(\mathbf{x}_g, \theta, w), \hat{\theta}, \hat{w}), \quad (4)$$

where  $\hat{\theta}$  and  $\hat{w}$  are the pose parameters and skinning weights, respectively, that account for the residual non-rigid motion. We compute the normals  $\mathbf{n}$  by taking the derivative of the deformed points  $\hat{\mathbf{x}}$  and  $\hat{\mathbf{x}}_g$ .

### 3.3 Learning 3D Dressed Human

In this subsection, we present our full 3D human recovery framework for monocular video. We start from describing the initialization of the parameters, followed by the optimization scheme.

#### 3.3.1 Input Initialization

As shown in Fig.2, given a monocular video, we obtain Densepose (Güler et al., 2018), surface nor-

mals and depth (Jafarian and Park, 2021), optical flow (Teed and Deng, 2020), and silhouette image (Lin et al., 2021) for each frame from off-the-shelf networks. We use FrankMocap (Rong et al., 2021) to initialize SMPL pose  $\Theta = \{\theta_1, \dots, \theta_n\}$  and shape  $B = \{\beta_1, \dots, \beta_n\}$  parameters, as well as camera viewpoints  $\mathbf{V} = \{v_1, \dots, v_n\}$  for a sequence of  $n$  frames. We average SMPL shape parameters  $B$  over the sequence and represent it in UV space as the initial base shape  $d_{base}$  for the person.

As shown in Fig. 3, Densepose only predicts UV for the naked SMPL. We extract features of both Densepose IUUV and RGB images using Principal Component Analysis (PCA) to segment the body parts for the entire region of the person, including clothing, in the image, and adopt a linear conversion to uniformly expand UV in each part. We use this extended IUUV together with depth prediction to initialize shape displacement  $\Delta d$  in the canonical UV space.

In order to separate illumination from reflectance in scenes for better rendering, we decompose each RGB image into albedo and shading images (Bell et al., 2014). Given extended IUUV and albedo images, we initialize the diffuse albedo UV map  $\rho$  by incorporating bi-linear interpolation. Given shading image and surface normals, we compute the light source direction vectors  $\mathbf{L} = \{l_1, \dots, l_n\}$  at each frame using a linear least-square solution:

$$l = (\mathbf{N}^T \mathbf{N})^{-1} \mathbf{N}^T \mathbf{S}, \quad (5)$$

where  $\mathbf{N} \in \mathbb{R}^{m \times 3}$  and  $\mathbf{S} \in \mathbb{R}^{m \times 1}$  are the normal matrix and the shading matrix with  $m$  sampled pixel points, respectively. Here we assume orthographic projection on the light source, and assume Lambertian reflection on the target surfaces.

### 3.3.2 Optimization

Given the initial parameters, we augment the shape displacement and texture UV maps to the naked SMPL (Sec. 3.1), and then transform it from the canonical space to the observation space using the deformer described in Sec. 3.2. We forward render its silhouette, normal, depth, densepose, and texture images with a differentiable renderer (Ravi et al., 2020). We define a rendering term  $\mathcal{L}_{ren}$  to enforce the consistency between the observation and the synthesized images:

$$\begin{aligned} \mathcal{L}_{ren}(\mathbf{V}, \mathbf{L}, \Theta, \hat{\Theta}, \hat{w}, \Delta d, \rho) = \\ \mathcal{L}_{sil} + \lambda_{tex} \mathcal{L}_{tex} + \lambda_{2D} \mathcal{L}_{2D} + \lambda_n \mathcal{L}_n + \lambda_d \mathcal{L}_d, \end{aligned} \quad (6)$$

where  $\mathcal{L}_{sil}$  and  $\mathcal{L}_{tex}$  are the silhouette loss and texture loss, respectively,  $\mathcal{L}_{2D}$  is the sum of Densepose reprojection loss, and  $\mathcal{L}_n$  and  $\mathcal{L}_d$  are used to ensure geometric consistency between predicted and synthesized

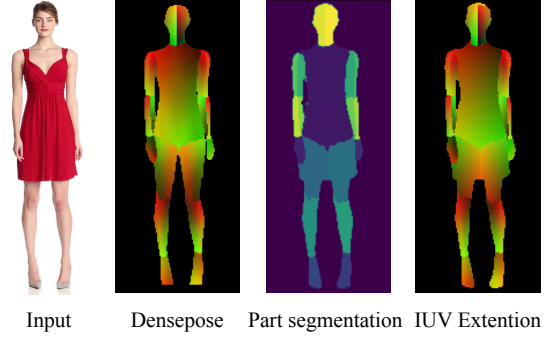


Figure 3: Initialization of extended IUUV.

geometry, respectively. As the visual structure is also important for reconstructing high fidelity results, we maximize the structural similarity by minimizing the dissimilarity:  $(1 - \text{MS-SSIM})/2$  (Wang et al., 2003), (Alldieck et al., 2019).  $\lambda_{tex}$ ,  $\lambda_{2D}$ ,  $\lambda_n$ , and  $\lambda_d$  are the weights that determine the relative importance of losses.

In order to better handle the invisible areas, we assume shape and texture are symmetric in each segment in the canonical UV space. For this, we enforce a symmetric structure constraint to the canonical shape displacement  $\Delta d$  and albedo  $\rho$  UV maps by minimizing:

$$\begin{aligned} \mathcal{L}_{sym}(\Delta d, \rho) = \sum_{i=1}^{10} \sum_{\mathbf{x} \in \Omega_i} \left\{ \|\Delta d(\mathbf{x}) - \Delta d(\mathbf{x}')\|^2 \right. \\ \left. + \lambda_\rho \|\rho(\mathbf{x}) - \rho(\mathbf{x}')\|^2 \right\}, \end{aligned} \quad (7)$$

where  $\Omega_i$  denotes the area of  $i^{\text{th}}$  segment in UV space,  $\lambda_\rho$  denotes the weight, and  $\mathbf{x}'$  is the flipped position of  $\mathbf{x}$  predefined for each segment.

To obtain a temporally coherent overall shape, as shown in Fig. 4, we define a regularization term  $\mathcal{L}_{reg}$  to enforce the temporal similarity of SMPL pose parameters  $\Theta$ , camera viewpoints  $\mathbf{V}$ , and light source direction vectors  $\mathbf{L}$  across frames:

$$\begin{aligned} \mathcal{L}_{reg}(\mathbf{V}, \mathbf{L}, \Theta) = \\ \sum_{i,j} \left\{ \epsilon_{ang}(l_i, l_j) + \epsilon_{rot}(\Theta_i, \Theta_j) + \lambda_v \|\mathbf{v}_i - \mathbf{v}_j\|^2 \right\}, \end{aligned} \quad (8)$$

where  $\epsilon_{ang}(\cdot)$  and  $\epsilon_{rot}(\cdot)$  are the angular error and the Riemannian distance (Moakher, 2002), respectively, and  $\lambda_v$  denotes the weight.

We further enforce the temporal consistency on the surface details by leveraging the optical flow (Teed and Deng, 2020) across frames:

$$\mathcal{L}_{tmp}(\Delta d, \Theta, \hat{\Theta}, \hat{w}) = \sum_{i,j} \sum_{\mathbf{p}_i} \|W_{i \rightarrow j}(\mathbf{p}_i) - \mathbf{p}_j\|^2, \quad (9)$$



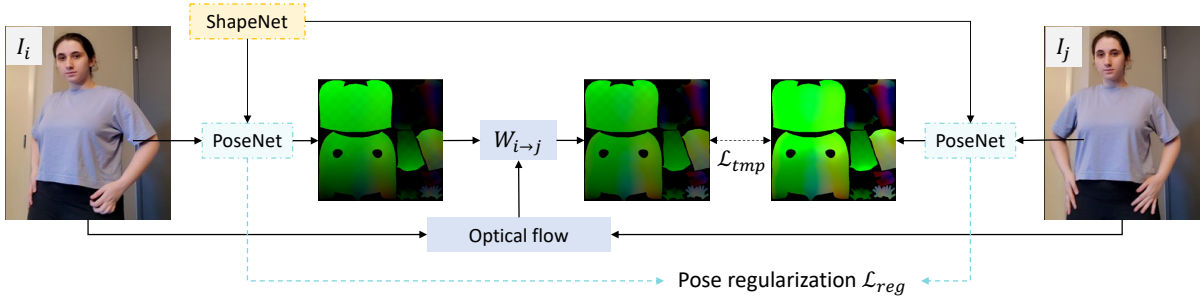


Figure 4: Temporal coherence. We apply temporal smoothness in the pose parameters, and apply temporal consistency in the rendered UV using optical flow.

Table 1: Quantitative Results. We report mean absolute error  $E_d$  in cm, mean angular error  $E_n$  in degree, image texture error  $\mathcal{L}_{tex}$  in RGB difference, and normal consistency error  $\mathcal{L}_n$  in degree, respectively (mean $\pm$ std).

Method	GT dress sequence		UBCFashion sequences	
	$E_d$	$E_n$	$\mathcal{L}_{tex}$	$\mathcal{L}_n$
Vid2Avatar (Guo et al., 2023) (w/ mask)	$1.08 \pm 0.47$	$50.52 \pm 3.41$	$27.55 \pm 2.36$	$4.68 \pm 1.90$
Ours	<b><math>1.04 \pm 0.44</math></b>	<b><math>18.93 \pm 15.83</math></b>	<b><math>12.52 \pm 12.41</math></b>	$7.55 \pm 14.03$

where  $\mathbf{p}_i$  is a pixel point of rendered UV in the  $i^{th}$  frame, and  $W_{i \rightarrow j}$  is the optical flow from frame  $i$  to frame  $j$  for mapping  $\mathbf{p}_i$  to  $\mathbf{p}_j$  in  $j^{th}$  frame.

Overall, the initial parameters can be further refined by alternating among refining camera viewpoints  $\mathbf{V}$  and SMPL and garment pose parameters:  $\{\Theta, \hat{\Theta}\}$ , shape displacement map  $\Delta d$ , and light source directions  $\mathbf{L}$  and albedo  $\rho$  until convergence:

$$\arg \min_{\xi_p \text{ or } \Delta d \text{ or } \xi_l} \mathcal{L}_{ren} + \lambda_{sym} \mathcal{L}_{sym} + \lambda_{reg} \mathcal{L}_{reg} + \lambda_{tmp} \mathcal{L}_{tmp} + \lambda_p \mathcal{L}_p, \quad (10)$$

where  $\xi_p = \{\mathbf{V}, \Theta, \hat{\Theta}, \hat{w}\}$  and  $\xi_l = \{\mathbf{L}, \rho\}$ ,  $\mathcal{L}_p$  is a L2 penalization term that prevents the pose parameters from deviating too much from the initialization,  $\lambda_{sym}$ ,  $\lambda_{reg}$ ,  $\lambda_{tmp}$  and  $\lambda_p$  denote loss weights.

### 3.4 Implementation Details

We regress a canonical UV representation consisting of geometry and texture for each subject. As shown in Fig. 2, our method consists of one U-Net for the canonical UV maps (*ShapeNet*) and one MLP for the motion parameters (*PoseNet*). We set the input RGB images to  $256 \times 256$  resolution, and set UV map to  $512 \times 512$  resolution to contain most details of the foreground while preventing from too much interpolation (Alldieck et al., 2019). The *ShapeNet* features each four convolution-batchnorm-ReLU down- and up-sampling layers. The *PoseNet* uses 4 layers of multi-layer perception with ReLU (Agarap, 2018) as the activation function after each layer. We use

Adam optimizer (Kingma and Ba, 2014) with batch size of 8 and learning rate of  $10^{-4}$ . We set  $\lambda_{tex} = 0.33$ ,  $\lambda_{2D} = 10^{-4}$ ,  $\lambda_n = 0.03$ ,  $\lambda_d = 0.02$ ,  $\lambda_{sym} = 10^{-2}$ ,  $\lambda_{reg} = 0.5$ ,  $\lambda_{tmp} = 10^{-4}$  and  $\lambda_p = 5 \times 10^{-4}$ . We use an NVIDIA V100 GPU and Intel(R) Xeon(R) CPU, and our model is implemented with Pytorch (Paszke et al., 2019).

## 4 EXPERIMENTS

We evaluate the effectiveness of our method on both synthetic and real data of people wearing loose clothing. We compare our method with baseline methods for full human 3D reconstruction from a monocular video.

**Datasets.** For the synthetic data, similar to Guo et al. (Guo et al., 2021), we generate a video sequence of SMPL model from the CMU motion capture database. Given shape and pose parameters of the human body model, we use ArcSim (Narain et al., 2012) to simulate the cloth motion of a dress from Berkeley Garment Library (Wang et al., 2011). For the real data, we use 3 fashion video sequences with subjects wearing loose clothing from *UBCFashion* dataset (Zablotskaia et al., 2019), as well as some Internet videos.

**Baseline Methods.** We quantitatively and qualitatively compare our method with state-of-the-art that focus on 3D reconstructing holistic human geometry from a single monocular video: Vid2Avatar (Guo

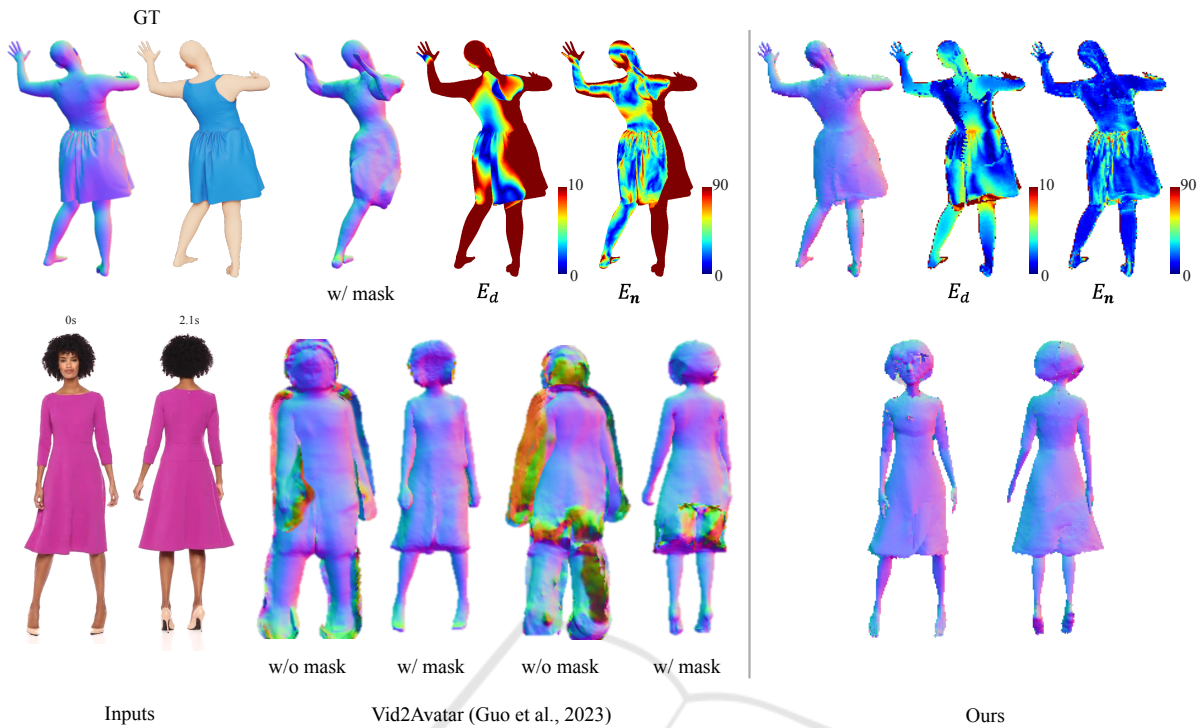


Figure 5: Qualitative comparison of 3D reconstruction on synthetic (top) and real (bottom) data. For the synthetic data, we show shape  $E_d$  and normal  $E_n$  error maps computed after aligning 3D reconstruction results with the ground truth. We can observe that our method achieves more accurate and detailed surface recovery.

Table 2: Ablation study on the simulated GT dress sequence. We report mean absolute error  $E_d$  (cm), mean angular error  $E_n$  (degree), image texture error  $E_{tex}$  (RGB), and temporal consistency error  $\mathcal{L}_{tmp}$  respectively (mean $\pm$ std).

Losses	$E_d$	$E_n$	$E_{tex}$	$\mathcal{L}_{tmp}$
$\mathcal{L}_{sil} + \mathcal{L}_{2D} + \mathcal{L}_{reg}$	$1.44 \pm 0.44$	$29.25 \pm 15.68$	-	$2.88 \pm 1.43$
$\mathcal{L}_{ren} \text{ (w/o } \mathcal{L}_{tex}) + \mathcal{L}_{reg}$	$1.41 \pm 0.46$	$19.22 \pm 15.58$	-	$2.60 \pm 1.22$
$\mathcal{L}_{ren} + \mathcal{L}_{sym} + \mathcal{L}_{reg}$	$1.42 \pm 0.45$	$19.08 \pm 15.89$	$12.10 \pm 14.30$	$2.36 \pm 0.96$
Full model	<b><math>1.04 \pm 0.44</math></b>	<b><math>18.93 \pm 15.83</math></b>	<b><math>12.02 \pm 13.12</math></b>	<b><math>1.97 \pm 0.86</math></b>

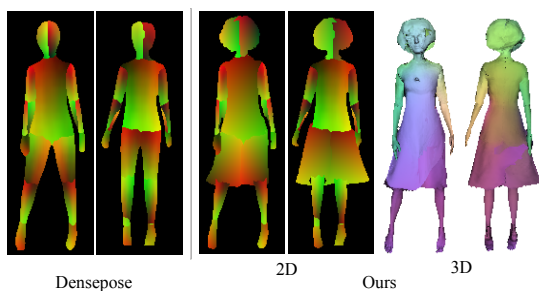


Figure 6: Qualitative UV results (same inputs as Fig. 5). We show Densepose UV (Güler et al., 2018) as well as ours UV in 2D and 3D space.

et al., 2023). Since our method can also generate reliable UV of dressed human observed in the video, we also qualitatively compare our method with the baseline method on human UV prediction: Densepose (Güler et al., 2018).

**Metrics.** For the synthetic data, we report the average geometry errors in posed space computed after aligning recovered 3D geometry with the ground truth in Table 1. For the real data, we warp canonical 3D model back to observation space and report the average error of two different rendering losses:  $\mathcal{L}_n$  and  $\mathcal{L}_{tex}$ , between the input images and the synthesized images (Table 1).

**Ablation Studies.** As reported in Table 2, we conduct an ablation study to analyze the impact of different losses. We can observe that our final model achieves the best performance in geometry and photometric errors, as well as temporal consistency error.

**Qualitative Results.** We evaluate our method qualitatively by visualizing the results of 3D geometry reconstruction to demonstrate the performance of the

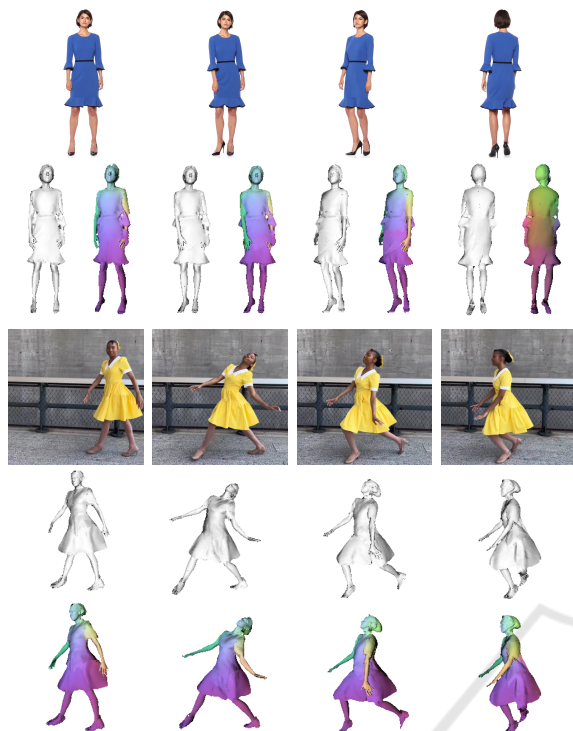


Figure 7: Results of our method on real videos. For each subject, we show the input frames, recovered geometry, and dense UV in 3D space from different viewpoints.

method in Fig. 5. For the synthetic data, we also show shape and normal error maps obtained after aligning the 3D reconstruction results with the ground truth. These results validate the accuracy of our method for recovering more accurate geometry of a person wearing loose clothing. We also show our UV recovery qualitatively in both 2D and 3D space in Fig. 6. This demonstrates the effectiveness of the method in tracking holistic surface correspondences. Fig. 7 shows more results on real videos.

**Garment Re-Texturing and/or Transfer.** As shown in Fig. 8, we take one result of our method and re-texture its garment by altering the albedo UV map using standard image editing techniques. Optionally, we can also modify the geometry UV map to apply garment transfer in 3D space in the posed space.

**Limitation.** As described in Sec. 3.2, we assume garment deformation closely follows the deformation of the body, our method cannot handle too complex garment non-rigid dynamics correctly. One of the possible future directions is to incorporate (Santesteban et al., 2021).



Figure 8: An example of re-texturing.

## 5 CONCLUSION

In this paper, we introduced a novel method for recovering a consistent, dense 3D geometry and appearance of a dressed person by observing it in a monocular video. We reconstruct the holistic 3D surface and texture represented in a canonical UV space. Our method jointly learns the shape displacement and albedo UV maps, as well as pose parameters with the differential neural rendering. In addition, we enhance the temporal coherence by utilizing a pose regularization term and the optical flow. Experimental results on real videos demonstrate the effectiveness of the method and the ability to perform dense 3D correspondence tracking of a person wearing loose clothing. We believe our work would expand the application of 3D human generation in a wide range of domains.

## REFERENCES

- Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.
- Alldieck, T., Pons-Moll, G., Theobalt, C., and Magnor, M. (2019). Tex2shape: Detailed full human body geometry from a single image. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 2293–2303.
- Alldieck, T., Zanfir, M., and Sminchisescu, C. (2022). Photorealistic monocular 3d reconstruction of humans wearing clothing. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*
- Angelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., and Davis, J. (2005). Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416.
- Bell, S., Bala, K., and Snavely, N. (2014). Intrinsic images in the wild. *ACM Trans. on Graphics (SIGGRAPH)*, 33(4).
- Blinn, J. F. and Newell, M. E. (1976). Texture and reflection in computer generated images. *Communications of the ACM*, 19(10):542–547.
- Dong, Z., Xu, K., Duan, Z., Bao, H., Xu, W., and Lau, R. (2022). Geometry-aware two-scale pifu represen-

- tation for human reconstruction. *Advances in Neural Information Processing Systems*, 35:31130–31144.
- Floater, M. S. (2003). Mean value coordinates. *Computer aided geometric design*, 20(1):19–27.
- Güler, R. A., Neverova, N., and Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 7297–7306.
- Guo, C., Jiang, T., Chen, X., Song, J., and Hilliges, O. (2023). Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*
- Guo, J., Li, J., Narain, R., and Park, H. S. (2021). Inverse simulation: Reconstructing dynamic geometry of clothed humans via optimal control. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*
- Hilton, A. and Starck, J. (2004). Multiple view reconstruction of people. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 357–364. IEEE.
- Jafarian, Y. and Park, H. S. (2021). Learning high fidelity depths of dressed humans by watching social media dance videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 12753–12762.
- Jiang, W., Yi, K. M., Samei, G., Tuzel, O., and Ranjan, A. (2022). Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer.
- Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., and Sheikh, Y. (2015). Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, pages 3334–3342.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Lin, S., Yang, L., Saleemi, I., and Sengupta, S. (2021). Robust high-resolution video matting with temporal guidance.
- Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16.
- Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., and Black, M. J. (2020). Learning to dress 3d people in generative clothing. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 6469–6478.
- Moakher, M. (2002). Means and averaging in the group of rotations. *SIAM J. Matrix Anal.*, 24(1):1–16.
- Narain, R., Samii, A., and O’Brien, J. F. (2012). Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG)*, 31(6):1–10.
- Newcombe, R. A., Fox, D., and Seitz, S. M. (2015). Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, pages 343–352.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.-Y., Johnson, J., and Gkioxari, G. (2020). Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*.
- Rong, Y., Shiratori, T., and Joo, H. (2021). Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 1749–1759.
- Santesteban, I., Thuerey, N., Otaduy, M. A., and Casas, D. (2021). Self-supervised collision handling via generative 3d garment models for virtual try-on. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, volume 2, page 3.
- Teed, Z. and Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer.
- Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., and Schmid, C. (2017). Learning from synthetic humans. In *CVPR*, pages 109–117.
- Wang, H., O’Brien, J. F., and Ramamoorthi, R. (2011). Data-driven elastic models for cloth: modeling and measurement. *ACM transactions on graphics (TOG)*, 30(4):1–12.
- Wang, L., Zhang, J., Liu, X., Zhao, F., Zhang, Y., Zhang, Y., Wu, M., Yu, J., and Xu, L. (2022). Fourier plenotrees for dynamic radiance field rendering in real-time. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 13524–13534.
- Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee.
- Weng, C.-Y., Curless, B., Srinivasan, P. P., Barron, J. T., and Kemelmacher-Shlizerman, I. (2022). Humanerf: Free-viewpoint rendering of moving people from monocular video. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 16210–16220.
- Xiu, Y., Yang, J., Cao, X., Tzionas, D., and Black, M. J. (2023). ECON: Explicit Clothed humans Optimized via Normal integration. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*
- Zablotskaia, P., Siarohin, A., Zhao, B., and Sigal, L. (2019). Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*.
- Zhao, F., Yang, W., Zhang, J., Lin, P., Zhang, Y., Yu, J., and Xu, L. (2022). Humannerf: Efficiently generated human radiance field from sparse inputs. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 7743–7753.