Supplementary Material: Animatable Neural Radiance Fields for Human Modeling

Sida Peng[1*] Junting Dong[1*] Qianqian Wang[2] Shangzhan Zhang[1]
Qing Shuai[1] Xiaowei Zhou[1] Hujun Bao[1†]
[1]Zhejiang University [2]Cornell University

In the supplementary material, we provide the derivation of transformation matrices, network architectures, details of training and test data, and 3D reconstruction results.

1. Derivation of transformation matrices

We represent the human skeleton as $(\mathbf{J}, \boldsymbol{\theta})$, where $\mathbf{J} \in \mathbb{R}^{K\times 3}$ denotes the joint locations of K joints and $\boldsymbol{\theta} \in \mathbb{R}^{3(K+1)\times 1} = [\boldsymbol{\omega}_0^T, ..., \boldsymbol{\omega}_K^T]$ denotes the $(K+1)$ relative rotation of body part with respect to its parent part in a kinematic tree using the axis-angle representation. Then, the transformation matrix of part k from canonical pose $\boldsymbol{\theta}_c$ to target pose $\boldsymbol{\theta}_t$ can be represented as

$$G_k = A_k(\mathbf{J}, \boldsymbol{\theta}_t) A_k(\mathbf{J}, \boldsymbol{\theta}_c)^{-1}, \tag{1}$$

$$A_k(\mathbf{J}, \boldsymbol{\theta}) = \prod_{i \in P(k)} \begin{bmatrix} R(\boldsymbol{\omega}_i) & \mathbf{j}_i \\ 0 & 1 \end{bmatrix}, \tag{2}$$

where $R(\boldsymbol{\omega}_i) \in \mathbb{R}^{3\times 3}$ is the converted rotation matrix of $\boldsymbol{\omega}_i$ via the Rodrigues formula, $\mathbf{j}_i$ is the i-th joint center, and $P(k)$ is the ordered set of parent joints of joint k. In practice, we adopt the SMPL skeleton [3], which has $K = 24$ parts, but this idea applies to other human skeletons [1, 2].

2. Network architectures

We present architectures of NeRF network and neural blend weight field network in Figures 1 and 2, respectively.

3. Training and test data

We show the detailed frame numbers for training and test of each subject in Table 1. Since the video length of each subject is different, we choose the appropriate number of frames (150∼300) to train the model and take remaining video frames for test.

4. 3D reconstruction

Figure 3 presents the reconstruction results in the canonical space in the first two columns. As described in the paper, we can use the learned blend weight field to animate the reconstructed geometry, which is also shown in Figure 3. We find that the original reconstruction tend to be rough, which may be caused by the inaccurate segmentation results. To

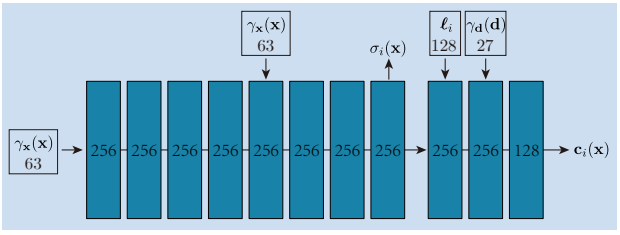

Figure 1. **Network architecture of the density and color fields.** The network is almost the same as the original NeRF, except that we introduce a per-frame latent code $\boldsymbol{\ell}_i$ to encode the state of human appearance in frame i. The number in each block means the dimension of the input.

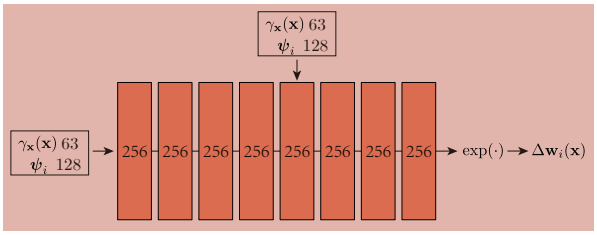

Figure 2. **Network architecture of the neural blend weight filed.** The network takes the positional encoding of the location $\gamma_{\mathbf{x}}(\mathbf{x})$ along with a per-frame latent code $\boldsymbol{\psi}_i$ and outputs the residual blend weight $\Delta\mathbf{w}_i(\mathbf{x})$ using exponential map. The network consists of 8 linear layers with ReLU activations and includes a skip connection on the fifth layer, which is similar to the density prediction module of the original NeRF. The number in each block means the dimension of the input.

	S1	S5	S6	S7	S8	S9	S11
training	150	250	150	300	250	260	200
test	49	127	83	200	87	133	82

Table 1. **Frame numbers for training and test of each subject.**

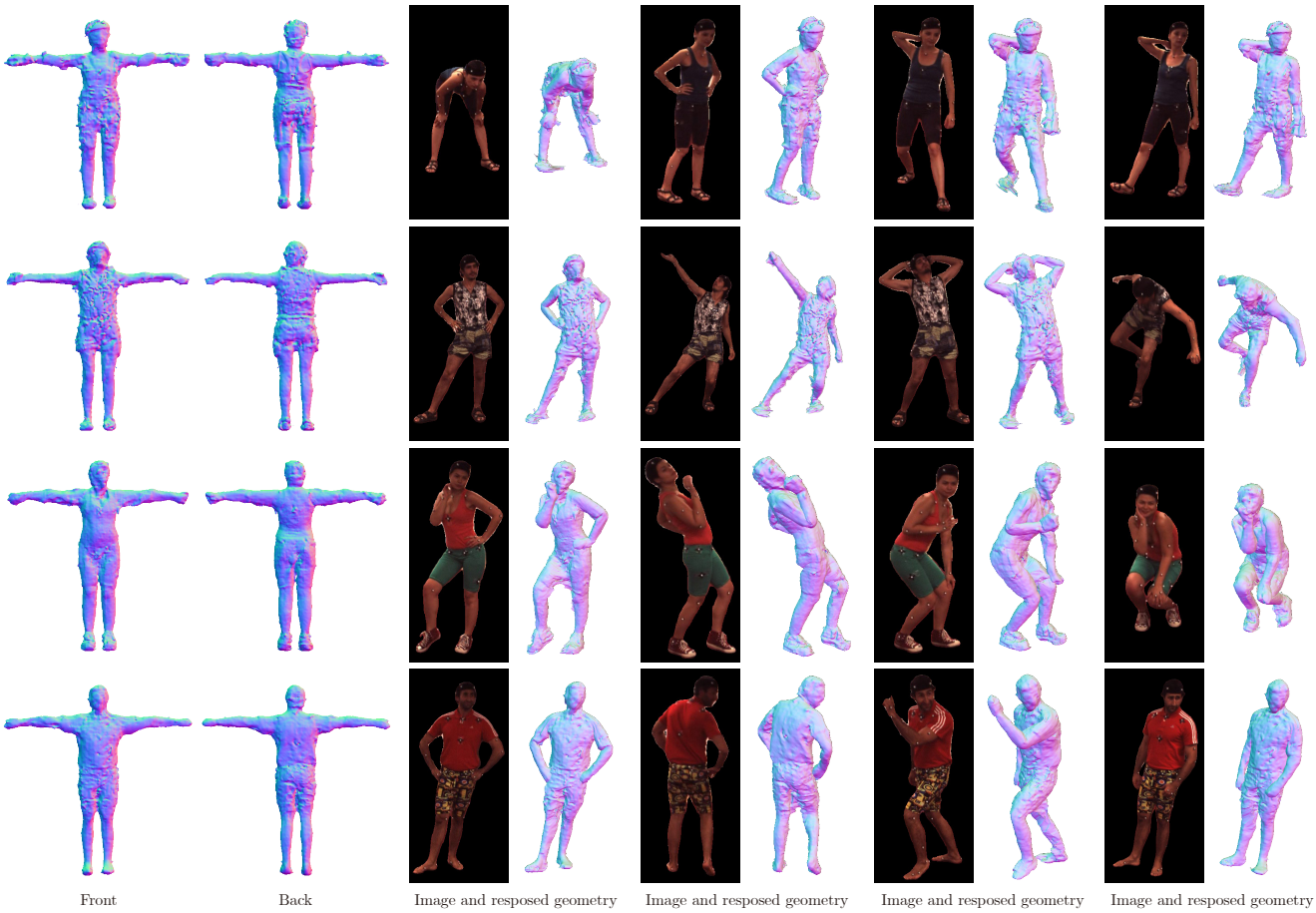

Figure 3. **Reconstructed geometries and reposed geometries.** The first two columns show the reconstructed geometries in the canonical space, which can be animated according to input human poses. We evaluate the mesh animation on test video frames.

solve this problem, we additionally apply Gaussian smoothing to the reconstructed geometry. We present more results in the supplementary video.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *PAMI*, 2018. 1

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2013. 1

[3] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 2015. 1