NECA: Neural Customizable Human Avatar — Supplementary Material

1. Implementation Details

1.1. Building Tangent Space

To construct the tangent space, we begin by computing the TBN (Tangent, Bitangent, Normal) matrix for the transformation. Consider a triangle with vertices $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, and edges $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$. The normal of the triangle is computed as the cross product of edges $\mathbf{e}_1$ and $\mathbf{e}_2$:

$$\mathbf{n}_f = \mathbf{e}_1 \times \mathbf{e}_2. \tag{1}$$

Now, focusing on vertex $\mathbf{v}_1$ (similar computations apply to other vertices), the normal of vertex $\mathbf{v}_1$ can be calculated by:

$$\mathbf{n}_1 = \frac{\sum_{m=1}^{N} \mathbf{n}_f^m}{\sum_{m=1}^{N} \|\mathbf{n}_f^m\|_2}, \tag{2}$$

where N is the number of triangles that vertex $\mathbf{v}_1$ belongs to, and $\mathbf{n}_f^m$ is the normal vector of the m-th triangle. The tangent and bitangent vectors of $\mathbf{v}_1$ can be calculated by solving a combination of linear equations. Given the texture coordinate (u_i^*, v_i^*) of vertex $\mathbf{v}_i$, we express the edges as linear combinations:

$$\begin{aligned} \mathbf{e}_1 &= \Delta u_1^* \mathbf{t}_1 + \Delta v_1^* \mathbf{b}_1, \\ \mathbf{e}_2 &= \Delta u_2^* \mathbf{t}_1 + \Delta v_2^* \mathbf{b}_1. \end{aligned} \tag{3}$$

This can also be written component-wise as:

$$\begin{aligned} (e_{1x}, e_{1y}, e_{1z}) &= \Delta u_1^* (t_{1x}, t_{1y}, t_{1z}) + \Delta v_1^* (b_{1x}, b_{1y}, b_{1z}), \\ (e_{2x}, e_{2y}, e_{2z}) &= \Delta u_2^* (t_{1x}, t_{1y}, t_{1z}) + \Delta v_2^* (b_{1x}, b_{1y}, b_{1z}), \end{aligned} \tag{4}$$

where Δu_1^* and Δv_1^* are differences in texture coordinates along edge $\mathbf{e}_1$, Δu_2^* and Δv_2^* are differences along edge $\mathbf{e}_2$, and $\mathbf{t}_1$ and $\mathbf{b}_1$ represent the tangent and bitangent vectors for vertex $\mathbf{v}_1$. These equations can be expressed as matrix multiplication:

$$\begin{bmatrix} e_{1x} & e_{1y} & e_{1z} \\ e_{2x} & e_{2y} & e_{2z} \end{bmatrix} = \begin{bmatrix} \Delta u_1^* & \Delta v_1^* \\ \Delta u_2^* & \Delta v_2^* \end{bmatrix} \cdot \begin{bmatrix} t_{1x} & t_{1y} & t_{1z} \\ b_{1x} & b_{1y} & b_{1z} \end{bmatrix}. \tag{5}$$

Solving for the tangent and bitangent vectors:

$$\begin{bmatrix} t_{1x} & t_{1y} & t_{1z} \\ b_{1x} & b_{1y} & b_{1z} \end{bmatrix} = \begin{bmatrix} \Delta u_1^* & \Delta v_1^* \\ \Delta u_2^* & \Delta v_2^* \end{bmatrix}^{-1} \cdot \begin{bmatrix} e_{1x} & e_{1y} & e_{1z} \\ e_{2x} & e_{2y} & e_{2z} \end{bmatrix}. \tag{6}$$

For an arbitrary surface point $\mathbf{x}_s$, its normal $\mathbf{n}_s$ can be computed by:

$$\mathbf{n}_s = \mathcal{B}_{u_s^*, v_s^*}(\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3), \tag{7}$$

where $\mathcal{B}$ denotes barycentric interpolation, (u_s^*, v_s^*) represents the UV coordinate of $\mathbf{x}_s$, and $\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3$ are the normals of the three vertices of the triangle that $\mathbf{x}_s$ falls into. The computation of tangent and bitangent vectors is similar.

1.2. Network Architecture

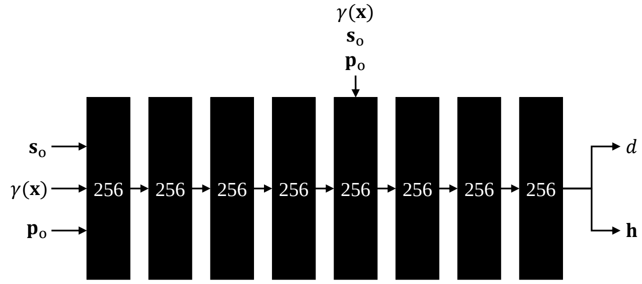


Figure 1. **SDF Network.** We show the architecture of our SDF network, which takes subject-level feature $\mathbf{s}_o$, canonical position $\mathbf{x}$, and pose-aware feature $\mathbf{p}_o$ as input, and outputs signed-distance d and latent feature vector $\mathbf{h}$. A skip connection that concatenates $\gamma(\mathbf{x})$, $\mathbf{s}_o$ and $\mathbf{p}_o$ to the fifth layer is employed. Except for the last layer, each layer outputs 256 dimension features with softplus activations.

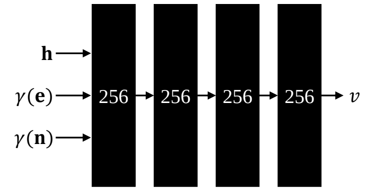


Figure 2. **Shadow Network.** We show the architecture of our shadow network, which takes latent feature $\mathbf{h}$, viewing direction $\mathbf{e}$ and normal $\mathbf{n}$ as input, and outputs shadow v. Besides the last layer, each layer outputs 256 dimension features with ReLU activations.

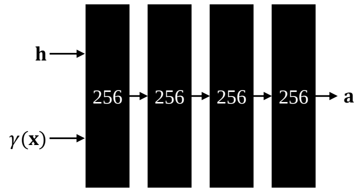


Figure 3. **Albedo Network.** We show the architecture of our albedo network, which takes latent feature $\mathbf{h}$ and canonical position $\mathbf{x}$ as input, and outputs albedo $\mathbf{a}$. Each layer outputs 256 dimension features with ReLU activations, except for the last layer.

The detailed architecture of different sub-networks in our framework are illustrated in Figs. 1 to 3. For canonical position $\mathbf{x}$, viewing direction $\mathbf{e}$, and normal $\mathbf{n}$, we employ positional encoding γ to enhance the ability of capturing high-frequency details.

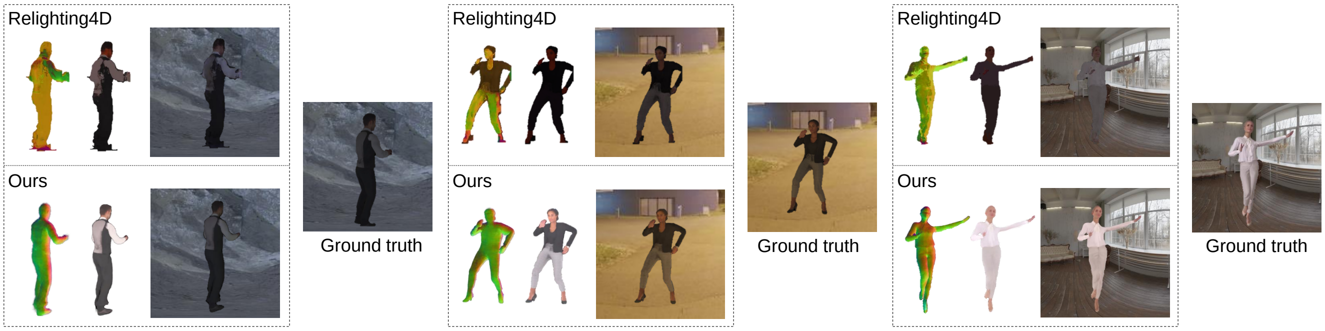


Figure 4. **Qualitative comparison of relighting on the synthetic dataset.** As can be seen, in contrast to the results of Relighting4D, our generated relighting results are visually closer to the ground truths. Best view in color.

Model	PSNR ↑	SSIM ↑	LPIPS ↓
Relighting4D [3]	21.9	0.811	0.201
Ours	**22.6**	**0.843**	**0.159**

Table 1. **Quantitative comparison of relighting under novel pose on the synthetic dataset.** Our method outperforms Relighting4D on the synthetic dataset with ground truth relighting results.

Model	Novel View			Novel Pose		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Rel. Pos.	28.0	0.955	0.052	26.0	0.936	0.067
Dir.	27.9	0.954	0.053	25.9	0.935	0.067
UVH	27.8	0.954	0.053	25.9	0.935	0.067
Ours	**28.3**	**0.957**	**0.051**	**26.3**	**0.940**	**0.066**

Table 2. **Ablation study on alternatives for our local tangent coordinate.** "Rel. Pos." means relative position between sampled point and nearest point, while "Dir." indicates direction and "UVH" represents barycentric coordinate and distance.

Num of R	Novel View			Novel Pose			Param.
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	
1	27.9	0.953	0.051	25.8	0.934	0.066	1.70M
12	28.0	0.954	0.051	25.9	0.935	0.066	5.82M
48	28.3	**0.957**	0.051	**26.3**	**0.940**	0.066	19.30M
64	**28.4**	**0.957**	0.051	**26.3**	**0.940**	0.066	25.30M

Table 3. **Ablation study on the number of R.** As shown, larger R corresponds to overall better performance on novel view and pose synthesis, but this trend becomes less obvious when $R \geq 48$. To trade off the performance and the model size, $R = 48$ is utilized in all our experiments.

	Novel View			Novel Pose		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NB [11]	19.3	0.889	0.129	17.4	0.863	0.151
AN [10]	17.8	0.875	0.154	16.7	0.855	0.164
ARAH [12]	19.5	0.893	0.124	17.8	0.868	0.144
PV [9]	20.2	0.903	0.105	18.0	0.870	0.121
Ours	**20.9**	**0.910**	**0.100**	**19.1**	**0.883**	**0.117**

Table 4. **More quantitative comparison on DeepCap and DynaCap datasets.**

	Novel View			Novel Pose		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
HN [13]	30.4	0.974	0.024	23.8	0.936	0.069
Ours	**30.9**	**0.978**	**0.023**	**30.9**	**0.977**	**0.023**

Table 5. **Quantitative comparison with HumanNeRF on subject "377" in ZJU-Mocap.**

2. Relighting Results on Synthetic Dataset

2.1. Dataset Details

To quantitatively evaluate the performance of our method on human relighting, we create a synthetic dataset with 12 videos using 3 publicly available 3D characters from [2], rigged with animations from [1]. The dataset is rendered under 4 different lighting conditions, including one natural sunlight and three HDRI maps. Each monocular video has about 200 frames, where only 30 frames rendered under natural sunlight are used for training, the rest of frames are used for evaluation. Following previous work, we employ [7] to estimate the SMPL and camera parameters.

2.2. Quantitative and Qualitative Results

Tab. 1 demonstrates that our method clearly outperforms Relighting4D [3] on the synthetic dataset. Visual comparison results are presented in Fig. 4, where our method exhibits superior performance on relighting, even in dealing with a challenging dataset with only 30 frames for training. In comparison, Relighting4D [3] struggles to produce physically convincing relighting renderings.

3. Additional Quantitative Results

We present additional quantitative comparison results on the DeepCap [4] and DynaCap [5] datasets in Tab. 4. Specifically, we evaluate the performance of ”lan” from DeepCap and ”vlad” from DynaCap. For each subject, we employ 300 frames for training and another 300 frames for evaluation. Notably, our method demonstrates superior performance over all baseline models by a significant margin.

Additionally, we conduct a comparison with the monocu-

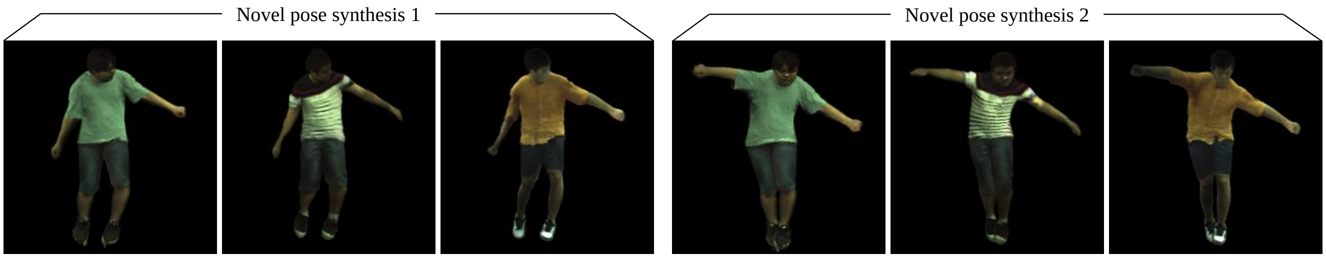


Figure 5. **More reposing results on ZJU-MoCap dataset.** We show two groups of novel pose synthesis results with poses from AIST++ [8].

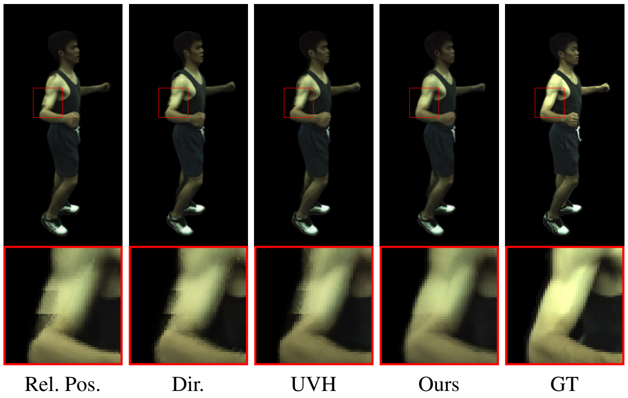

Rel. Pos. Dir. UVH Ours GT

Figure 6. **Ablation study on different local coordinate representations.** "Rel. Pos." means relative position between sampled point and nearest point, "Dir." indicates direction and "UVH" represents barycentric coordinate and distance.

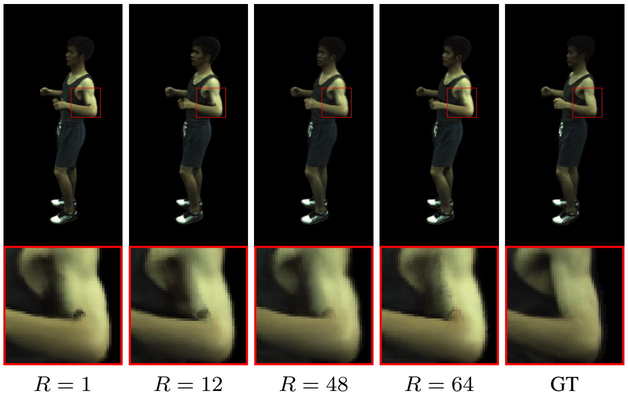

$R = 1$ $R = 12$ $R = 48$ $R = 64$ GT

Figure 7. **Impact of the number of components R in CP decomposition.**

lar method HumanNeRF [13] in Tab. 5. Our results indicate that our approach achieves superior performance, particularly in the context of novel pose synthesis.

4. Additional Ablation Studies

4.1. Ablation Study on Local Coordinate

As described in our paper, we introduce a novel local coordinate defined in tangent space. To validate the effectiveness of our local coordinate, we compare it with three alternatives. One common choice is the relative position. Denoting the sampled point as $\mathbf{x}_o$ and the nearest point on SMPL as $\mathbf{x}_s^*$, the relative position $\mathbf{c}_{pos}$ is expressed as $\mathbf{c}_{pos} = \mathbf{x}_o - \mathbf{x}_s^*$. Another alternative is the direction $\mathbf{c}_{dir}$, defined as the normalization of the relative position: $\mathbf{c}_{dir} = \frac{\mathbf{c}_{pos}}{\|\mathbf{c}_{pos}\|_2}$. The third option is UVH, which includes barycentric coordinate and distance: $\mathbf{c}_{uvh} = (u, v, h)$, where (u, v) denotes the barycentric coordinate, h represents the distance between $\mathbf{x}_o$ and $\mathbf{x}_s^*$. To examine the performance of these alternatives, we conduct evaluation on the "377" subject in ZJU-MoCap [11], and provide the quantitative results in Tab. 2. Comparing the results, we can see that our proposed local coordinate achieves the best results on all three metrics, manifesting its effectiveness and superiority. In addition, we present the qualitative comparison results in Fig. 6, where our local coordinate in tangent space exhibits better alignment effect.

4.2. Ablation Study on Component Number R

To assess the influence of different component numbers R in the CP decomposition, we conduct experiments on the subject "377" from ZJU-MoCap. The numerical results for novel view and pose synthesis are summarized in Tab. 3. As shown, higher component number R corresponds to better performance, but incurs larger model size. We opt for $R = 48$ in our method, for striking a balance between the performance and the model size. The visual comparison results on different R are shown in Fig. 7, where we can see that the generated renderings with larger R are visually better with less visual artifacts. Note that, the renderings with $R = 48$ and $R = 64$ are visually similar, showing that $R = 64$ would not induce noticeable performance increase compared to $R = 48$.

5. More Customization Results

We also showcase additional results in Figs. 5 and 8 to 11 to further highlight the versatility of our method in various customization tasks. For a comprehensive overview, we encourage reviewers to refer to the supplementary video.

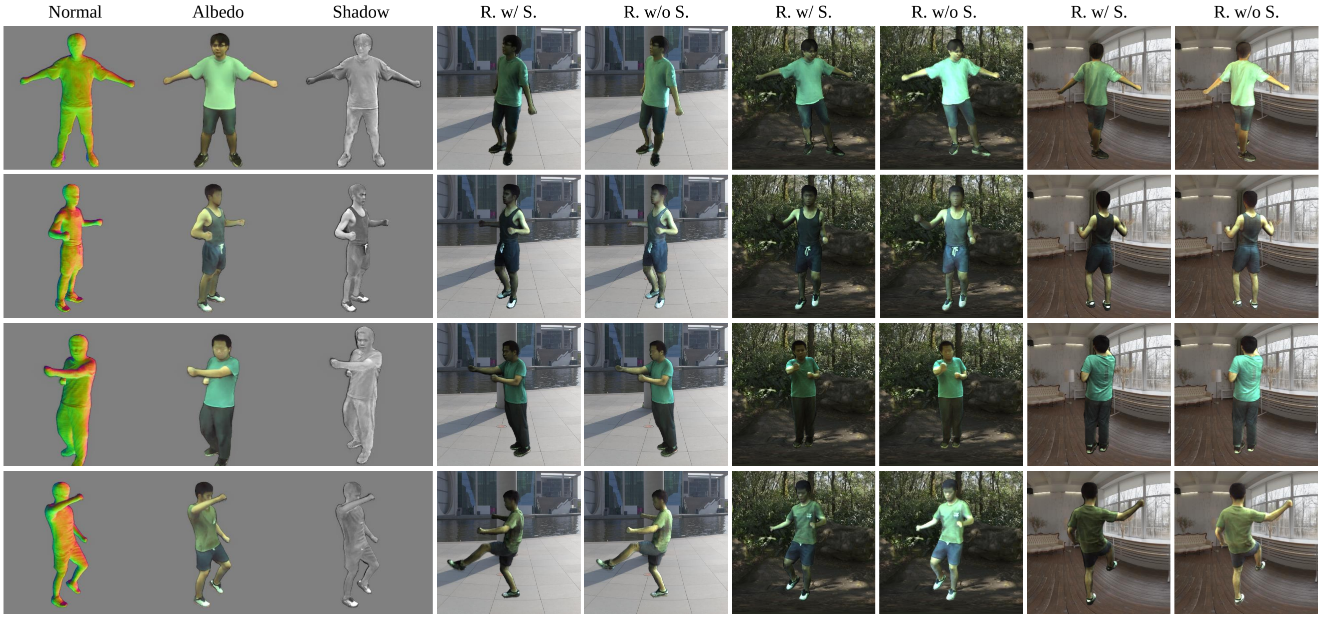


Figure 8. **More relighting results on ZJU-MoCap dataset.** “R. w/ S.” and “R. w/o S.” refer to results with and without shadows.

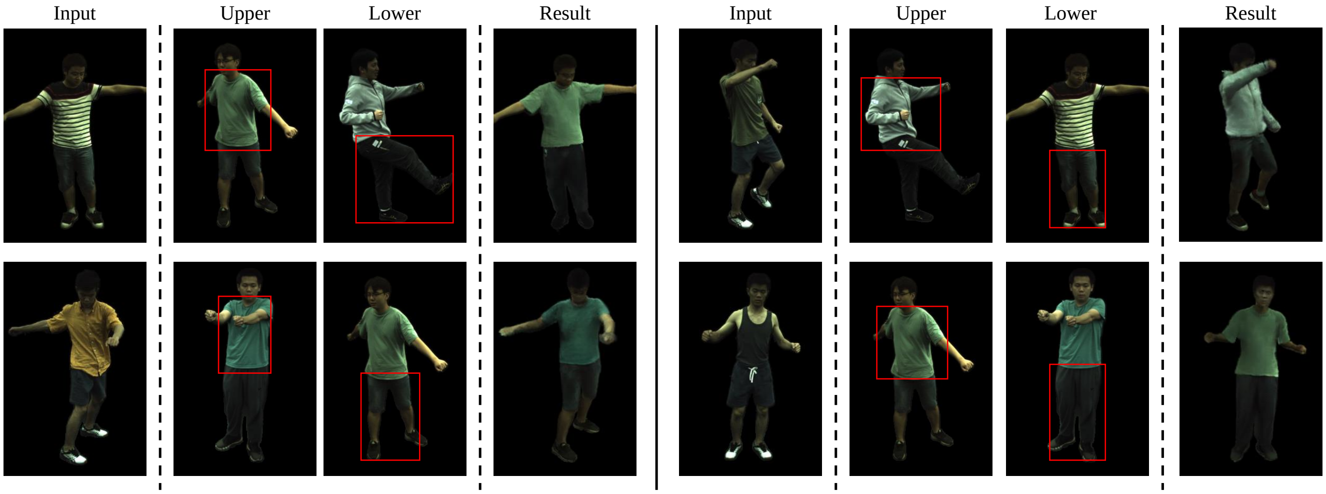


Figure 9. **More retexturing results on ZJU-MoCap dataset.**

6. Ethics Statement

The datasets utilized in our research are sourced from publicly available repositories, including ZJU-MoCap [11], NeuMan [6], DynaCap [5] and DeepCap [4]. Additionally, we incorporate a dataset generated using 3D characters from RenderPeople [2]. The meticulous collection of these datasets aligns with ethical guidelines and principles.

Given the capabilities of our work in generating realistic animated humans, it is crucial to acknowledge potential ethical considerations. NECA has the capacity to depict humans in various scenarios, including wearing different clothing with realistic lighting and shadow effects. While this technology offers creative possibilities, it also raises concerns about the potential misuse of generating misleading or fabricated videos of real individuals, contributing to negative social impacts. To address this, we emphasize the responsible and ethical use of our technology. Moreover, we recognize the environmental impact of the computational resources required by our method, which could contribute to increased carbon emissions. In response, we commit to releasing our pretrained weights to promote computational efficiency and reduce the environmental footprint associated with using our approach. This proactive measure aims to balance technological advancements with ethical considerations.

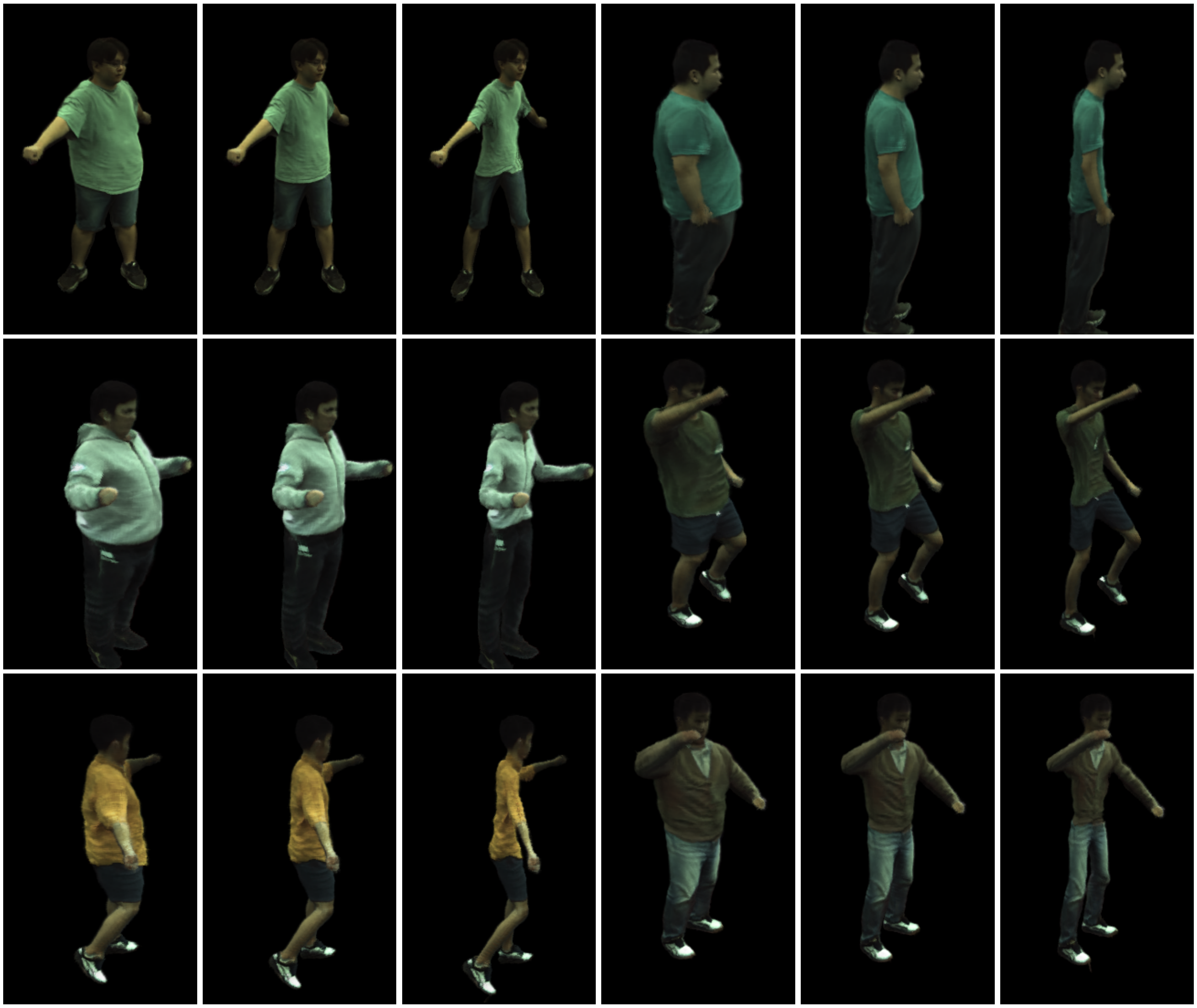
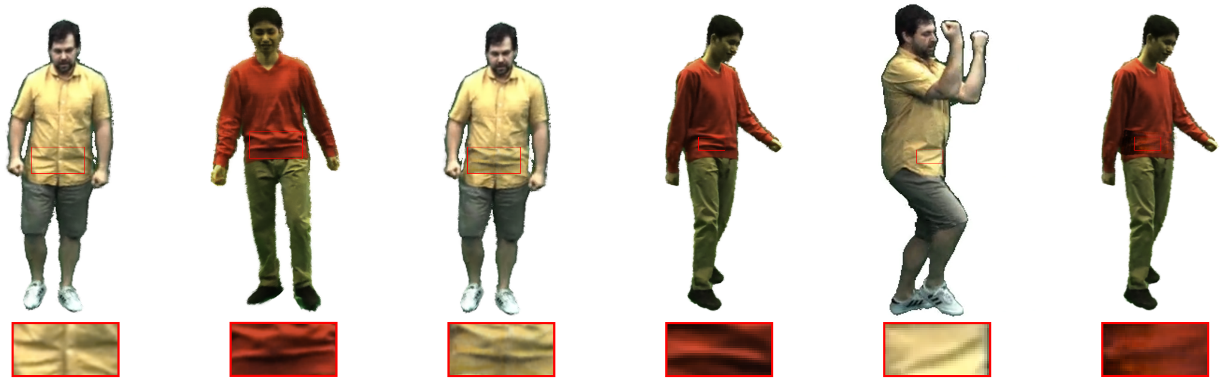


Figure 10. **More shape editing results on ZJU-MoCap dataset.**



Input Reference Result Input Reference Result

Figure 11. **More shadow transfer results on DynaCap [5] and DeepCap [4] datasets.**

References

[1] Mixamo. `https://www.mixamo.com/`. 2

[2] Renderpeople. `https://renderpeople.com/`. 2, 4

[3] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *ECCV*, 2022. 2

[4] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 2, 4, 5

[5] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Trans. Graphics*, 40(4), 2021. 2, 4, 5

[6] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, 2022. 4

[7] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021. 2

[8] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation. In *ICCV*, 2021. 3

[9] Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In *SIGGRAPH*, 2023. 2

[10] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 2

[11] 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. In *CVPR*, 2021. 2, 3, 4

[12] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *ECCV*, 2022. 2

[13] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 2, 3