Supplementary Material: HumanNeRF

A. Derivation of Motion Bases

We describe how we derive the rotation and translation, $\{R_i,\ \mathbf{t}_i\}$, to map from bone coordinates in observation space to coordinates in canonical space (Section 3 on "skeletal motion").

We define body pose $\mathbf{p} = (J, \Omega)$, where $J = \{j_i\}$ includes K joint locations and $\Omega = \{\boldsymbol{\omega}_i\}$ defines local joint rotations using axis-angle representations $\in \mathfrak{so}(3)$. Given a predefined canonical pose $\mathbf{p}_c = (J^c, \Omega^c)$ and an observed pose $\mathbf{p} = (J, \Omega)$, the observation-to-canonical transformation M of body part k is:

$$M_k(\mathbf{p}_c, \mathbf{p}) = \prod_{i \in \tau(k)} \begin{bmatrix} \exp(\boldsymbol{\omega}_i^c) & j_i^c \\ 0 & 1 \end{bmatrix} \left\{ \prod_{i \in \tau(k)} \begin{bmatrix} \exp(\boldsymbol{\omega}_i) & j_i \\ 0 & 1 \end{bmatrix} \right\}^{-1}, \tag{16}$$

where $\exp(\boldsymbol{\omega}) \in SO(3)$ is a 3×3 rotation matrix computed by taking the exponential of $\boldsymbol{\omega}$ (i.e., applying Rodrigues' rotation formula), and $\tau(k)$ is the ordered set of parents of joint K in the kinematic tree.

The rotation and translation, R_k and t_k, for body part k is can then be extracted from M_k:

$$\begin{bmatrix} R_k & \mathbf{t}_k \\ 0 & 1 \end{bmatrix} = M_k(\mathbf{p}_c, \mathbf{p}). \tag{17}$$

B. Network Architecture

Figures 9-12 show the network design for the canonical MLP, the non-rigid motion MLP, the pose correction MLP, and the deep network generating the canonical motion weight volume.

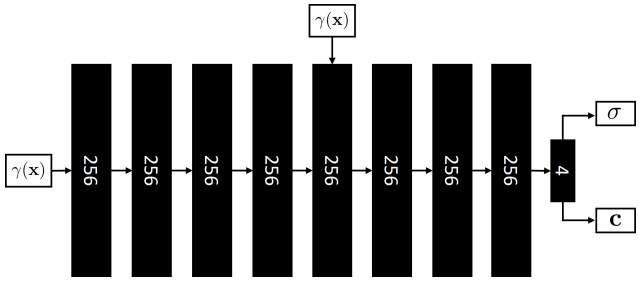

Figure 9. Canonical MLP visualization. Following NeRF [2], we use an 8-layer MLP with width=256, taking as input positional encoding γ of position $\mathbf{x}$ and producing color $\mathbf{c}$ and density σ. A skip connection that concatenates $\gamma(\mathbf{x})$ to the fifth layer is applied. We adopt ReLU activation after each fully connected layer, except for the one generating color $\mathbf{c}$ where we use *sigmoid*.

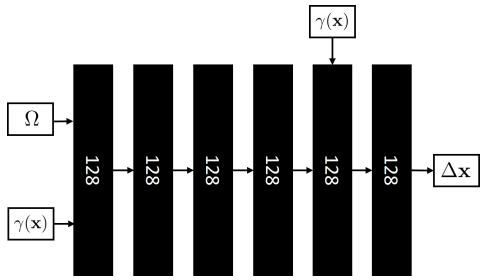

Figure 10. Non-rigid motion MLP visualization. We choose a 6-layer MLP (width=128) that takes as input the body pose, specifically, joint rotations Ω, and positional encoding, $\gamma(\mathbf{x})$, and predicts the offset $\Delta\mathbf{x}$. We use a skip connection for the positional encoding at the fifth layer. Additionally, we remove the rotation vector of global orientation from joint angles Ω and only uses the remainder as MLP input.

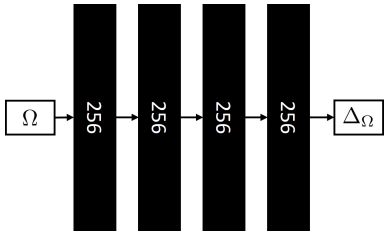

Figure 11. Pose correction MLP visualization. A 4-layer MLP with width 256 that takes joint angles Ω is used for refining initial poses. Like the non-rigid motion MLP, we take all joints except for root joint (i.e., body orientation) into account and optimize them accordingly.

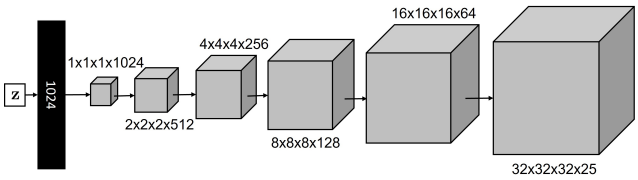

Figure 12. Network for generating the motion weight volume. The network begins with a fully-connected layer that transforms the (random, constant) latent code $\mathbf{z}$ and reshapes it to a $1 \times 1 \times 1 \times 1024$ grid. Subsequently, it is concatenated with 5 transposed convolutions, increasing volume size while decreasing the number of channels, and finally, produces a volume of size $32 \times 32 \times 32 \times 25$. LeakyReLU is applied after MLP and transposed convolution layers. The size of the latent code $\mathbf{z}$ is 256.

C. Motion Field Decomposition

We decompose a motion field into skeletal rigid motion and non-rigid motion. We tested several different formulations for the decomposition. Specifically, starting from a point $\mathbf{x}$ in observation space, we considered three po-

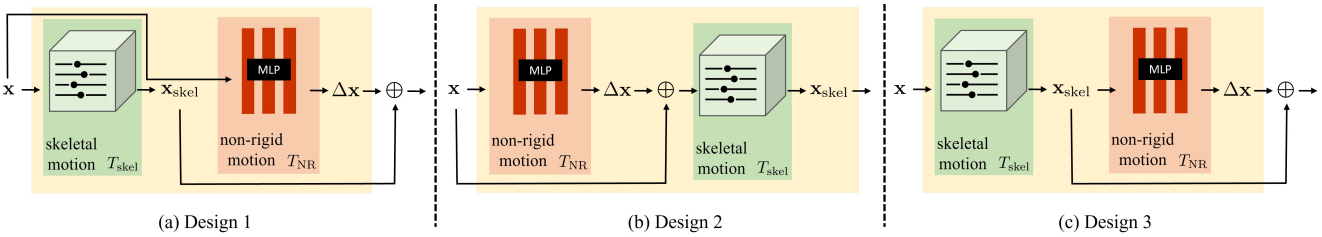

Figure 13. The three proposed designs of motion decomposition. We choose design 3 (c) as a result of best quality of novel view synthesis, shown in Fig. 14.

tential decompositions. (To simplify notation and improve readability below, we omit body pose $\mathbf{p}$, which would otherwise always appear as the second argument to each of $T, T_{\text{skel}}, T_{\text{NR}}$.)

(1) Both T_{skel} and T_{NR} conditioned on an observed point position $\mathbf{x}$, illustrated in Fig. 13-(a):

$$T(\mathbf{x}) = T_{\text{skel}}(\mathbf{x}) + T_{\text{NR}}(\mathbf{x}) \tag{18}$$

(2) T_{NR} conditioned on $\mathbf{x}$, but T_{skel} conditioned on position adjusted by non-rigid motion, $\mathbf{x} + T_{\text{NR}}(\mathbf{x})$, illustrated in Fig. 13-(b):

$$T(\mathbf{x}) = T_{\text{skel}}(\mathbf{x} + T_{\text{NR}}(\mathbf{x})) \tag{19}$$

(3) T_{skel} conditioned on $\mathbf{x}$ and T_{NR} conditioned on the position $T_{\text{skel}}(\mathbf{x})$ warped by skeletal rigid motion T_{skel}, illustrated in Fig. 13-(c):

$$T(\mathbf{x}) = T_{\text{skel}}(\mathbf{x}) + T_{\text{NR}}(T_{\text{skel}}(\mathbf{x})) \tag{20}$$

We conducted experiments on the PeopleSnapshot dataset [1], and used 64 samples per ray for quick evaluation. As shown in Fig. 14, deforming $\mathbf{x}$ by T_{skel} and then conditioning T_{NR} on that motion (design 3, or Eq. 20) produces the best quality for novel view synthesis. The result of this experiment explains our final choice of motion decomposition.

D. Additional Implementation Details

There are several small but important implementation details that contribute to best results. We describe them below.

Optimizing $\Delta\mathbf{W_c}$: Our method solves for W_c to determine skeletal rigid motion. In practice, we ask a deep network to generate ΔW_c instead, the difference between W_c and the logarithm of W_g. W_g consists of an ellipsoidal Gaussian around each body bone, given by the canonical T-pose, that specifies approximate body part regions in the canonical space. W_c is then computed as:

$$W_c = softmax(\Delta W_c + \log(W_g)), \tag{21}$$

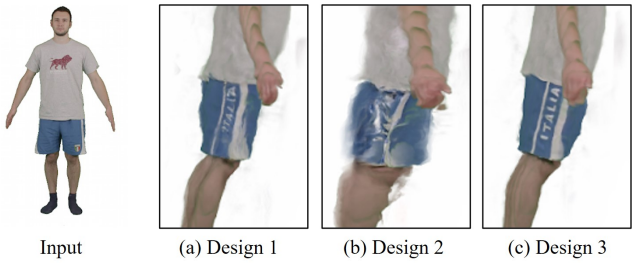

Figure 14. The experimental result of novel view synthesis on the three proposed motion decompositions, illustrated in Fig. 13. Design 3 (c) leads to best alignment, the approach we ultimately adopted. In this experiment, we used 64 samples per ray for quick evaluation, introducing color artifacts on the arms not present when using the sampling described in the paper.

where the background weight in W_g is set to one minus the sum of all the bone weights. We apply the logarithm to W_g, to compensate the exponential function in *softmax*.

Representation of global body orientation: Global subject orientation can be represented as body rotation or, equivalently, camera rotation. We choose to rotate the camera in order to keep the estimated bounding box when subject orientation changes. Specifically, we uses axis-aligned bounding boxes because for ease of implementation; however, the box will be different for the same pose but rotated global body orientation. This undesirable effect can be avoided if we instead describe changes of global body orientation as camera rotations.

Random background: During optimization, we randomly assign a solid background color to the rendering and to the input image to facilitate separation of foreground and background.

MLP initialization: We initialize the weights of the last layer of the non-rigid motion MLP and pose correction MLP to small values, $\mathcal{U}(-10^{-5}, 10^{-5})$, i.e., initializing the offset to be close to zero and the pose refinement rotation matrices each near the identity.

Importance ray sampling: We sample more rays for the foreground subject, indicated by the segmentation masks. Specifically, we enforce random ray sampling with proba-

	Subject **313**			Subject **315**			Subject **390**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Neural Body [4]	29.417	0.9635	57.24	26.93	0.9597	55.97	29.57	0.9609	52.12
Ours	29.421	0.9672	29.54	26.65	0.9636	33.76	30.52	0.9682	33.88

Table 4. Additional quantitative comparison on ZJU-MoCap dataset. We color cells having the best metric value. LPIPS* = LPIPS $\times 10^3$.

bility 0.8 for foreground subject pixels and 0.2 for the background region.

E. More Results

E.1. Additional Results

We conduct an additional experiment on the remaining three subjects (313, 315, 390) in ZJU-MoCap dataset. The results are shown in Table 4. Consistent with the results in the main paper, our method outperforms NeuralBody, particularly under the perceptual metric LPIPS. Fig. 16 shows visual comparisons. Our method substantially captures the appearance details for unseen regions while Neural Body produces blurry results.

E.2. Ablation Study on Sequence Length

To understand how our method performs on different sequence lengths, we evaluate it on the sequences that vary in the number of frames but are sampled from the same video. Specifically, we take subject 392 from ZJU-MoCap dataset and use images captured from "camera 1" temporally sub-sampled at rates of 1, 2, 5, 10, and 30, yielding five training sequences containing 556, 228, 112, 56, and 19 frames respectively. For evaluation, we use the same motion sequence temporally sub-sampled by 30 but captured from the other 22 cameras not seen in the training. We use the same hyperparameters and training iterations throughout. The results are shown in Table 5.

	PSNR ↑	SSIM ↑	LPIPS* ↓
556 frames	31.04	0.9705	32.12
228 frames	30.84	0.9701	31.78
112 frames	31.01	0.9703	32.75
56 frames	30.90	0.9693	35.45
19 frames	30.51	0.9655	45.17

Table 5. Ablation study on sequence length. We color cells with best metric values. LPIPS* = LPIPS $\times 10^3$.

As expected, using more frames leads to better quality; however the improvement is not obvious when the frame number is over a threshold (in this case, 112 frames). We speculate that diversity of body poses is a more significant factor in reconstruction quality than number of frames.

E.3. Optimized Canonical Appearance

Fig. 17 shows the recovered appearance for the predefined T-pose on the ZJU-MoCap [4] dataset; the results for self-captured and YouTube videos are shown in Fig. 18.

E.4. Limitations

We provide two visual examples of our method's limitations in Fig. 15. Pose correction may fail if the video frame contains artifacts, e.g., strong motion blur, as shown in (a) and (b). Non-rigid motion was not fully recovered in (c) and (d), as the movement of the jacket depended on the temporal dynamics of subject motion.

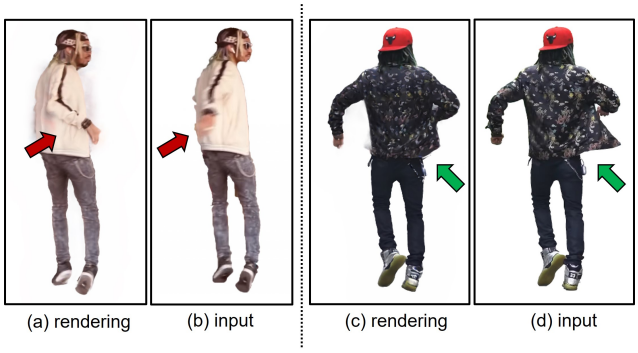

Figure 15. Visual examples of limitations. Pose correction may fail (a and b) and non-rigid clothes motion was not able to be fully recovered (c and d).

F. Societal Impact

In this work we aim to faithfully reproduce motion sequences performed by a person with the capability of rendering unseen views. Therefore applying the technology to create false depictions, e.g., re-animating the subject in novel poses, was not considered as a potential application. Nevertheless, the public deployment of the technology should still be done with care, e.g., by reminding audiences that imagery is computer-generated when adjusting the viewpoint. In addition, the high computation requirement of the algorithm may lead to increased carbon emissions. We hope the methods that accelerate training of neural graphics primitives (e.g., [3]) will help reduce computation and thus the environmental impact. Finally, our method will be made available to the public for counter-measure analysis and computation reduction.

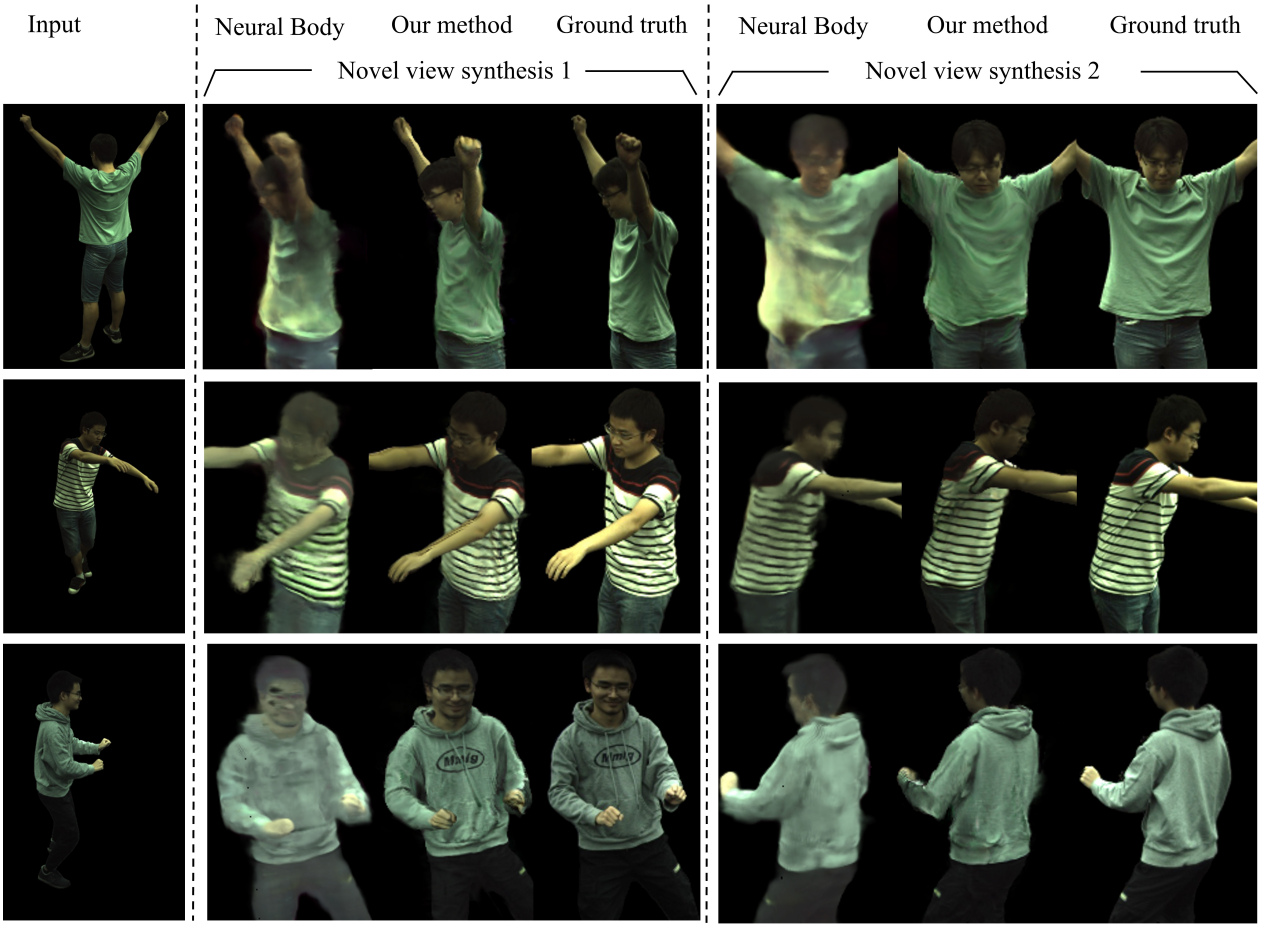

Figure 16. Qualitative comparison on the remaining subjects in ZJU-MoCap dataset.

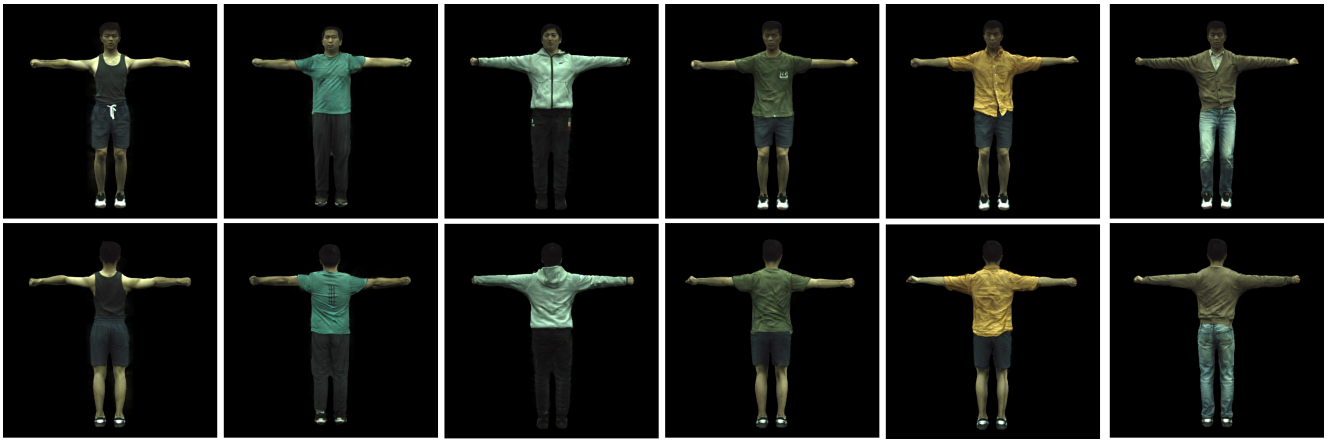

Figure 17. Optimized canonical appearance on ZJU-MoCap dataset.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[2] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *ECCV*, 2020. 1

[3] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multires-

Figure 18. Optimized canonical appearance for self-captured videos (first two columns) and YouTube videos (right three).

olution hash encoding. *arXiv:2201.05989*, Jan. 2022. 3

[4] 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. *CVPR*, 2021. 3