Supplemental Document
Mirror-Aware Neural Humans

In addition to the results in the main paper, we provide more qualitative results and implementation details. The supplemental video shows the results in motion.

1. Additional qualitative 3D Reconstruction Results

Figure 1 demonstrates the capability of Mirror-Aware Neural Humans (Step 2 output) to reconstruct challenging poses where existing 3D pose estimators are prone to fail. The lying down pose is representative both for strong self occlusions, which necessitates the second mirror view or other means of multi-view recording, and for extreme vertical poses, which goes against the bias towards standing poses learning-based approaches have.

2. Novel View and Pose Synthesis

Figure 2 shows novel view synthesis results that reveal an accurate 3D reconstruction of the monocular input images. Notably, the model-free nature enables reconstructing unique shapes such as the beard (first two rows), loose clothing (central two rows), and texture details such as logos (last two rows).

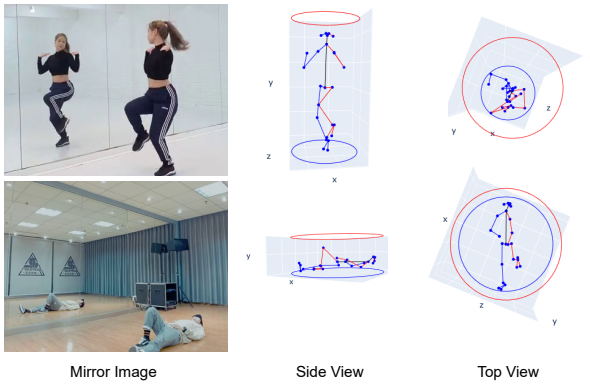

Figure 1. **Mirror 3D reconstruction**. Our mirror approach successfully reconstructs challenging virtual poses in the mirror from the side and top view from scratch, including an extreme case of lying on the ground.

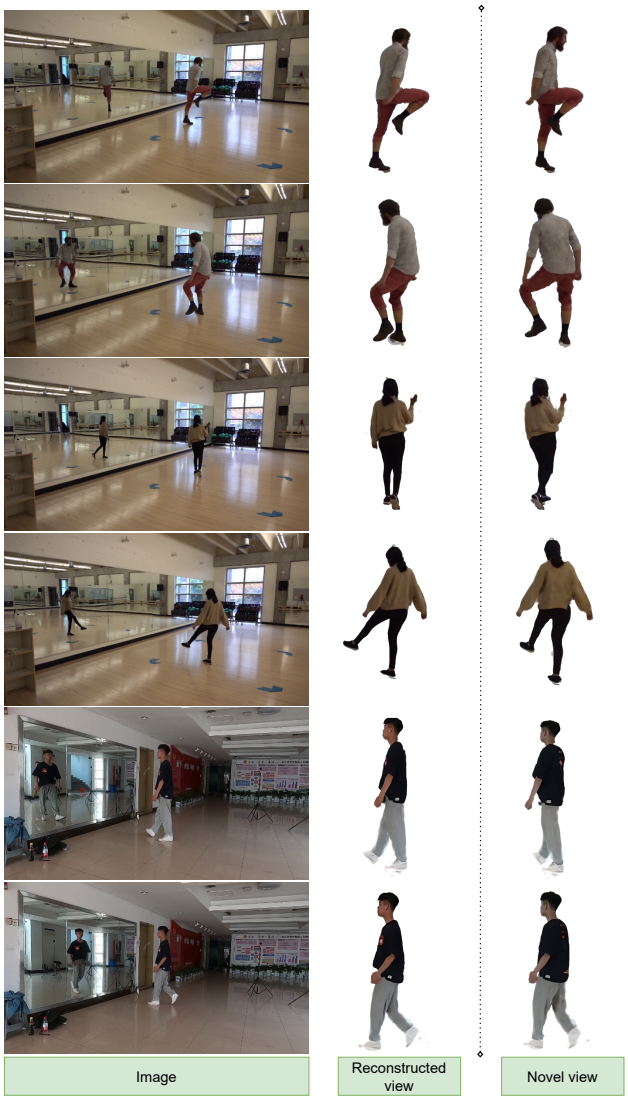

Figure 2. **Novel view synthesis results**. The learned reconstruction is a proper 3D model that can be rotated to novel views.

Figure 3 simulate one of the learned volumetric body models in an unseen mirror arrangement, requiring both

Figure 3. **Mirror Retargeting**. Given a source body (left), Mirror-Aware Neural Humans successfully simulate the mirror arrangement in an unseen setting (right).

novel view (mirror) and novel pose (transferred from source). This highlights the effectiveness of our automatic mirror normal estimation (discussed in the main paper).

3. Effect of Regularization Constraints

Figure 5 Left measures the importance of the smoothness term on our 3D reconstruction objective and shows its robustness to different weights. The smoothness term on orientation with a weight of 1.0 minimizes the 3D reconstruction error (measured in P-MPJPE). Figure 5 Right carries out the same ablation on the body joint location smoothness term, with a scale of 0.22, leading to the best results on the validation sequence.

4. Mirror Matrix Derivation

We provide further details on the mirror matrix derivation. Following [4, 6], we define a mirror operation $\mathbf{A}$ that mirrors points across the mirror plane π. This operation includes a rotation and reflection, and can be described using the plane equation,

$$n_x x + n_y y + n_z z + d = 0 \tag{1}$$

where $d = -\mathbf{m} \cdot \mathbf{n}_m$ is the distance of the mirror plane to the camera origin. Generally, $\mathbf{m}$ can be any point on the plane that defines the plane's location, and in this case, the mirror location is chosen as the point between the real and mirrored person. $\mathbf{n}_m$ is the normal of the mirror that defines the orientation of the plane. Hence, we can derive the virtual camera through a mirror matrix $\mathbf{A}$ that reflects point through the plane in Eq. 1 using $\mathbf{A} = \mathbf{I} - 2\mathbf{n}_m\mathbf{n}_m^T$ [2] where $\mathbf{I}$ is a 3×3 identity matrix. The mirror matrix $\mathbf{A}$ can be expanded as,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 2 \cdot \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} \cdot \begin{bmatrix} n_x & n_y & n_z \end{bmatrix} \text{ or, } \tag{2}$$

$$\mathbf{A} = \begin{bmatrix} (1-2n_x^2) & -2n_y n_x & -2n_z n_x \\ -2n_y n_x & (1-2n_y^2) & -2n_y n_z \\ -2n_z n_x & -2n_y n_z & (1-2n_z^2) \end{bmatrix}, \tag{3}$$

which can be fully expressed as a 4×4 affine transformation matrix such that,

$$\mathbf{A} = \begin{bmatrix} (1-2n_x^2) & -2n_y n_x & -2n_z n_x & -2n_x d \\ -2n_y n_x & (1-2n_y^2) & -2n_y n_z & -2n_y d \\ -2n_z n_x & -2n_y n_z & (1-2n_z^2) & -2n_z d \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{4}$$

with $\mathbf{n}_m = [n_x, n_y, n_z]$ the mirror normal and d is the distance between the camera and mirror (both quantities are from Step 1 in the main document. By defining the real camera to be at the origin pointing along the z-axis, $\mathbf{A}$ maps points from the real to the virtual camera using the following,

$$\begin{bmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{5}$$

Note that the orientation of the virtual camera switches to a left-hand coordinate system (LH) [5, 6] based on the reflection component in $\bar{\mathbf{R}}$.

> IOU score	No of Overlap Frames		Frame Ratio	
	C6	C7	C6	C7
5%	402	652	0.25%	0.42%
7%	322	569	0.20%	0.36%
10%	229	450	0.14%	0.29%
20%	63	72	0.04%	0.05%

Table 1. **Number of candidate frames and ratio** in cameras 6 and 7 automatically identified as overlapping cases based on their bounding box IOU score, thereby selecting the option with the highest confidence (>20%).

5. Separating Real and Mirror Poses

Since human pose detectors are not completely accurate, we drop frames where the pose prediction is below half of the height of a valid pose visually. On the MirrorHuman-eval dataset, we use the available GT focal length and remove detections on bystanders and correct ambiguous real-to-mirror associations using the distance to the 2D GT annotation. Note that this ensures that we work with automatic detections as available in practice, containing slight inaccuracies, while leaving complete mis-detections or assignments out of the equation to aid a fair comparison to previous methods using manually annotated ground truth 2D annotations. Note that this correction is only applied to the MirrorHuman-eval dataset and only to very few frames and cameras; only for camera 2 (2 mismatches out of 1571 frames) and camera 5 (4 mismatches out of 1377 frames).

To analyze the effect of frames with occlusion compared to frames without, we measure the IOU between the bounding boxes for the virtual and real person. We select the one with a high confidence above 20% and quantify the amount of occluded frames in Table 1. Only very few frames show occlusions but these are important to address.

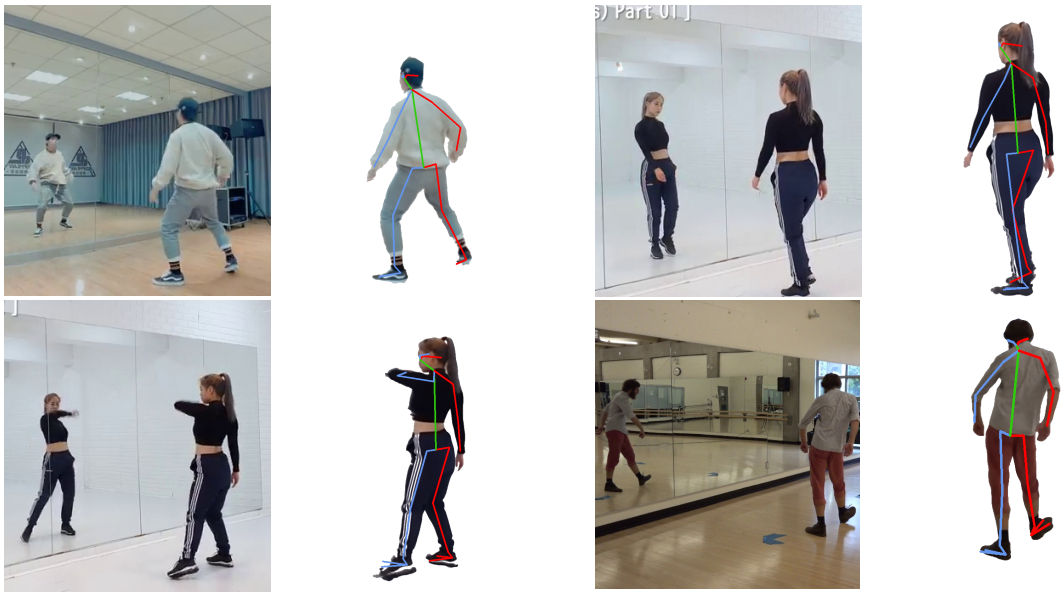

Figure 4. **Volumetric body model and 3D skeleton overlay**. Besides the ones provided in the main paper, we here show additional qualitative results on new actors and scenes. The skeleton overlay reveals how precise the fit is, even in unusual poses.

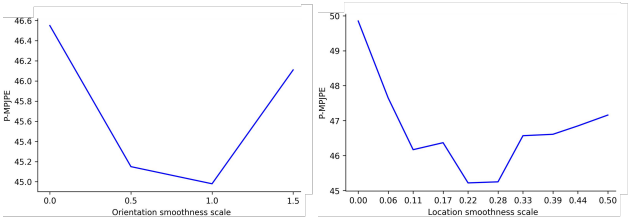

Figure 5. Ablation study on smoothness terms, all values in P-MPJPE. **Left:** Effect of the angular smoothness weight **Right:** Effect of the positional smoothness weight.

6. Optimization and Training Details

In Step 1, we solve for the focal length and initial ground plane by solving a system of equations using Singular Value Decomposition (SVD) for 20K RANSAC iterations. In Step 2, we jointly optimize for the joint rotations $\boldsymbol{\theta}$, pelvis position $\mathbf{p}_{\text{pelvis}}$, bone lengths $\boldsymbol{\ell}$, and joint rotations $\boldsymbol{\theta}$ (as explained in the main document) using the Adam optimizer [3] with a learning rate of 1×10^{-1} and without any weight decay. For a sequence of 2000 frames, we found our optimization to converge after 2K iterations in 2 hours on average with 16 Intel Core i7 CPU cores. Since our skeleton model has only few parameters we run Step 1 and Step 2 on the CPU.

In Step 3, we use the estimated pose and the bone rotations to learn $G(\theta)$. For occlusion handling, we use 20% threshold for automatic overlap area detection. Similar to [7], we sample a batch of 3072 rays in each iteration, and attribute 30% of these rays to occlusion areas if occlusion is detected. We train the volumetric model for 200K iterations and jointly refine the 3D pose. As in [7], we finetune the remaining body details for 145K, keeping pose fixed. As [8] does not refine pose, we train for a maximum of 300K iterations to learn the body model with the pose fixed. Optionally, we refine the 3D pose estimates from Step 2 with [7] to provide better initialization for [8]. The training is done on 4 Nvidia-Tesla V100 GPU and takes a 4-6 days. For example, for a sequence of 1620 training frames trained with occlusion handling, *Mirror DANBO w/ Occlusion* takes 30 hours for 300K iterations and *Mirror A-NeRF w/ Occlusion*, including pose refinement and finetuning, takes 130 hours for up to 400K iterations.

7. Camera Calibration Details

Hyperparameter selection. Hyperparameters, including the number of RANSAC steps and the angle threshold for up-right poses used in the camera calibration are determined on camera 2 and 3 of the MirrorHuman-eval set (our validation set).

Derivation. Our single view calibration is based on [1]. For completeness, we provide an overview of the relevant steps below. The single view calibration method is based on the well-established direct linear transform (DLT) [9] method to solve projective relations. We first write our constraints as a linear system of equations that is solved using Singular Value Decomposition (SVD), up to the unknown scale factor arising from the projection. To reach the form $\mathbf{M}\mathbf{x} = 0$, we take the cross product on both sides of Eq. 1

and Eq. 2 from the main paper and subtract the 2 results to derive

$$\mathbf{q}_{\text{neck}} \times K(\mathbf{p}_{\text{ankle}} + \mathbf{n} \cdot h) - \mathbf{q}_{\text{ankle}} \times K(\mathbf{p}_{\text{ankle}}) = 0, \quad (6)$$

with h the person height, $\mathbf{n}$ the normal direction, and $\mathbf{q}_{\text{neck}}$ and $\mathbf{q}_{\text{ankle}}$ the neck and ankle positions. In the following we subscript variables with an x, y, z to indicate the x,y,z-coordinates and with a number $1, 2, 3$ to refer to different person locations.

In matrix form, using $\Delta\mathbf{q}_x = \mathbf{q}_x^{\text{neck}} - \mathbf{q}_x^{\text{ankle}}$, $\Delta\mathbf{q}_y = \mathbf{q}_y^{\text{neck}} - \mathbf{q}_y^{\text{ankle}}$, and z to represent the unknown depth of the ankle, Eq. 6 can be expressed as

$$\begin{pmatrix} 0 & -1 & \mathbf{q}_y^{\text{neck}} & 0 & -1 & \Delta\mathbf{q}_y \\ 1 & 0 & -\mathbf{q}_x^{\text{neck}} & 1 & 0 & -\Delta\mathbf{q}_x \end{pmatrix} \begin{pmatrix} f\mathbf{n}_x \\ f\mathbf{n}_y \\ \mathbf{n}_z \\ \mathbf{n}_z\mathbf{o}_x \\ \mathbf{n}_z\mathbf{o}_y \\ z/h \end{pmatrix} = 0, \quad (7)$$

where f is the focal length and $\mathbf{o}$ the principal point of the camera intrinsics $\mathbf{K}$. By using at least three 2D neck $\mathbf{q}_{\text{neck}}$ and ankle $\mathbf{q}_{\text{ankle}}$ detections, we form the constraint matrix

$$\mathbf{D} = \begin{pmatrix} 0 & -1 & \mathbf{q}_{y1}^{\text{neck}} & \Delta\mathbf{q}_{y1} & 0 & 0 \\ 1 & 0 & -\mathbf{q}_{x1}^{\text{neck}} & -\Delta\mathbf{q}_{x1} & 0 & 0 \\ 0 & -1 & \mathbf{q}_{y2}^{\text{neck}} & 0 & \Delta\mathbf{q}_{y2} & 0 \\ 1 & 0 & -\mathbf{q}_{x2}^{\text{neck}} & 0 & -\Delta\mathbf{q}_{x2} & 0 \\ 0 & -1 & \mathbf{q}_{y3}^{\text{neck}} & 0 & 0 & \Delta\mathbf{q}_{y3} \\ 1 & 0 & -\mathbf{q}_{x3}^{\text{neck}} & 0 & 0 & -\Delta\mathbf{q}_{x3} \end{pmatrix} \quad (8)$$

that gives the system of equations

$$\mathbf{D} \begin{pmatrix} f\mathbf{n}_x + \mathbf{n}_z\mathbf{o}_x \\ f\mathbf{n}_y + \mathbf{n}_z\mathbf{o}_y \\ \mathbf{n}_z \\ z_1/h \\ z_2/h \\ z_3/h \end{pmatrix} = 0. \quad (9)$$

We solve Eq. 9 using SVD. Having more than three ankles and necks results in an over-determined system, for which we can find a least-squares solution.

Ground normal extraction. Since Eq. 9 is a 6×6 system with rank five, any solution we find is unique up to a scalar. In order to determine $\mathbf{n}$ from the SVD or least-squares solution, we use the fact that the normal vector is perpendicular to any vector formed by a pair of ankles. Using

$$\begin{bmatrix} \bar{\mathbf{n}_x} \\ \bar{\mathbf{n}_y} \\ \bar{\mathbf{n}_z} \\ \bar{z_1} \\ \bar{z_2} \\ \bar{z_3} \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{n}_x + \mathbf{n}_z\mathbf{o}_x/f \\ \mathbf{n}_y + \mathbf{n}_z\mathbf{o}_y/f \\ \mathbf{n}_z/f \\ z_1/(hf) \\ z_2/(hf) \\ z_3/(hf) \end{bmatrix}, \quad (10)$$

If we do not have a given focal length, we can derive the equation for the focal length,

$$f = \sqrt{\frac{(-(\bar{\mathbf{n}}_x - \bar{\mathbf{n}}_z\mathbf{o}_x)\bar{\mathbf{q}}_{\text{x}} - (\bar{\mathbf{n}}_y - \bar{n}_z\mathbf{o}_y)\bar{\mathbf{q}}_{\text{y}})}{(\bar{\mathbf{n}}_z(\bar{z}_1 - \bar{z}_2))}}, \quad (11)$$

with

$$\bar{\mathbf{q}}_{\text{x}} = ((\mathbf{q}_{x1}^{\text{ankle}} - \mathbf{o}_x)\bar{z}_1 - (\mathbf{q}_{x2}^{\text{ankle}} - \mathbf{o}_x)\bar{z}_2)p \quad (12)$$

and

$$\bar{\mathbf{q}}_{\text{y}} = ((\mathbf{q}_{y1}^{\text{ankle}} - \mathbf{o}_y)\bar{z}_1 - (\mathbf{q}_{y2}^{\text{ankle}} - \mathbf{o}_y)\bar{z}_2). \quad (13)$$

Either the estimated focal lengths or a given focal length enables us to recover $\lambda\mathbf{n}$ and $\lambda(z_1, z_2, z_3)$. To remove λ, we divide both vectors by the L_2 norm of $\lambda\mathbf{n}$, giving us a unique $\mathbf{n}$ of length one and ankle depths z_1, z_2, and z_3.

Using the normal vector $\mathbf{n}$ and the known depths z_1, z_2, and z_3, we recover the orientation and position of the ground plane and subsequently estimate the mirror plane from a 3D ankle position of a person and its mirror image, as explained in the main document.

References

[1] Anonymous. CasCalib: Cascaded Calibration for Motion Capture from Sparse Unsynchronized Cameras. *Supplemental document*, 2023. 3

[2] Alston S Householder. Unitary triangularization of a nonsymmetric matrix. *Journal of the ACM (JACM)*, 5(4):339–342, 1958. 2

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[4] Emőd Kovács. Rotation about an arbitrary axis and reflection through an arbitrary plane. In *Annales Mathematicae et Informaticae*, pages 175–186, 2012. 2

[5] Rui Rodrigues, Joao P Barreto, and Urbano Nunes. Camera pose estimation using images of planar mirror reflections. In *European Conference on Computer Vision*, pages 382–395. Springer, 2010. 2

[6] Katie Schwertz. Field guide to optomechanical design and analysis. Society of Photo-Optical Instrumentation Engineers, 2012. 2

[7] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems*, 2021. 3

[8] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. DANBO: Disentangled articulated neural body representations via graph neural networks. *arXiv preprint arXiv:2205.01666*, 2022. 3

[9] Ivan E Sutherland. Three-dimensional data input by tablet. *Proceedings of the IEEE*, 62(4):453–461, 1974. 3