Supplementary Material:
Reconstructing 3D Human Pose by Watching Humans in the Mirror

1. Estimating camera intrinsic parameters

From the projective geometry [1], we know that it is possible to calibrate the camera intrinsic parameters K from a single image.

Two orthogonal vanishing points: Suppose the camera has zero skew and square pixels, and the principal point is in the image center. K can be computed via two orthogonal vanishing points $\boldsymbol{v}_0 = [v_0^x, v_0^y, 1]^T$ and $\boldsymbol{v}_1 = [v_1^x, v_1^y, 1]^T$. The specific process is as follows:

According to the assumption, K has the following form:

$$K = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where f is the focal length. Let $\boldsymbol{\omega} = (KK^T)^{-1}$. Then,

$$\boldsymbol{\omega} = \begin{bmatrix} 1/f^2 & 0 & 0 \\ 0 & 1/f^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

As we have stated in the main manuscript, the ray through the camera center with direction $\boldsymbol{d}$ has the following relation with K and its vanishing point $\boldsymbol{v}$:

$$\boldsymbol{d} = K^{-1}\boldsymbol{v}. \tag{3}$$

We can obtain the angle between two rays through law of cosines:

$$\cos\theta = \frac{\boldsymbol{d}_0^T \cdot \boldsymbol{d}_1}{||\boldsymbol{d}_0|| \cdot ||\boldsymbol{d}_1||} = \frac{\boldsymbol{v}_0^T \boldsymbol{\omega} \boldsymbol{v}_1}{\sqrt{\boldsymbol{v}_0^T \boldsymbol{w} \boldsymbol{v}_0} \cdot \sqrt{\boldsymbol{v}_1^T \boldsymbol{w} \boldsymbol{v}_1}}. \tag{4}$$

For two vanishing points corresponding to orthogonal lines, we have:

$$\boldsymbol{v}_0^T \boldsymbol{\omega} \boldsymbol{v}_1 = 0. \tag{5}$$

Therefore, an analytical solution of f is as follows:

$$f = \sqrt{-(v_0^x v_1^x + v_0^y v_1^y)}. \tag{6}$$

Three orthogonal vanishing points: If we acquire three orthogonal vanishing points, the assumption about the principal point can be relaxed.

$$K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{7}$$

In this case, $\boldsymbol{\omega}$ can be represented as follows:

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_1 & 0 & \omega_2 \\ 0 & \omega_1 & \omega_3. \\ \omega_2 & \omega_3 & \omega_4 \end{bmatrix}. \tag{8}$$

Each pair of vanishing points $\boldsymbol{v}_i$, $\boldsymbol{v}_j$ can produce a linear equation $\boldsymbol{v}_i^T \boldsymbol{\omega} \boldsymbol{v}_j = 0$. Therefore, combining three pairs of vanishing points together forms the following equation:

$$\boldsymbol{A} \cdot \tilde{\omega} = \boldsymbol{0}, \tag{9}$$

where $\boldsymbol{A}$ is a 3×4 matrix and $\tilde{\omega} = [\omega_1, \omega_2, \omega_3, \omega_4]$.

$\tilde{\omega}$ is the null vector of $\boldsymbol{A}$. After obtaining $\tilde{\omega}$, K can be computed via Cholesky decomposition followed by matrix inversion.

2. Details of our multi-view system

The evaluation set contains several sequences with each around one minute long capturing a 1-2 person scene using six cameras (GoPro). The synchronization between the six cameras is done by a Wi-Fi Remote. Note that the mirror creates six virtual cameras, which are also available to use. Specifically, we regard the mirrored person as another view of the real person. The transformation between the real camera and the virtual camera is easy to obtain since the mirror is calibrated. Therefore, we have twelve cameras with known camera parameters to reconstruct 3D keypoints of the real person, which are used as ground-truth. 2D bounding boxes, 2D keypoints, and the correspondences across different views are annotated manually to ensure the accuracy. After obtaining the 3D keypoints of the real person by using triangulation, we can calculate the 3D keypoints of the corresponding mirrored person easily through the mirror geometry. The reconstruction results of our multi-view system are shown in the supplementary video.

3. Details of our Mirrored-Human dataset

Data collection: We collected the videos from online video websites like YouTube, bilibili and images from Google images. Various keywords were searched to guarantee the diversity, including *'mirror dance tutorial', 'pamela workout', 'install a large mirror',* etc. To ensure reconstruction quality, we selected the videos that most parts of the bodies are visible both inside and outside the mirror.

Annotations: The following points and edges are annotated manually:

- **2D joint positions** for all sampled frames in the videos as the ground truth.
- **Mirror edges** for the first frame of each video with a static camera, for all frames with a moving camera.

The following parameters are computed automatically:

- **Vanishing point** v_0 is calculated via 2D joint positions in each frame while v_1 is calculated via mirror edges.
- **Camera intrinsic** K (if unknown) is calculated via vanishing points v_0 and v_1.
- **Mirror normal** is calculated given v_0 and K. Note that if the camera is static, using all 2D joints in the video to estimate the mirror normal would be better.
- **Human model parameters** such as SMPL, SMPL-H and SMPL-X.

4. More results

More qualitative results can be found in our supplementary video. We show our reconstruction results for both images and videos. A temporal smoothing constraint that uses an L_1 penalty on the differences between two consecutive frames is added for the video input.

References

[1] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision.* Cambridge university press, 2003. 1