Supplementary Material
HUMBI: A Large Multiview Dataset of Human Body Expressions

*Zhixuan Yu† *Jae Shin Yoon† In Kyu Lee† Prashanth Venkatesh†
Jaesik Park‡ Jihun Yu♯ Hyun Soo Park†
†University of Minnesota ‡POSTECH ♯BinaryVR
{yu000064, jsyoon, leex7424, venka220, hspark}@umn.edu
jaesik.park@postech.ac.kr, jihun.yu@binaryvr.com

This supplementary material provides additional details of HUMBI.

A. Multi-camera System

We design a unique multi-camera system that was deployed in public events including Minnesota State Fair and James Ford Bell Museum of Natural History at the University of Minnesota. There are 772 subjects captured by 107 GoPro HD cameras recording at 60Hz.

Hardware The capture stage is made of a re-configurable dodecagon frame with 3.5 m diameter and 2.5 m height using T-slot structural framing (80/20 Inc.). The stage is encircled by 107 GoPro HD cameras (38 HERO 5 BLACK Edition and 69 HERO 3+ Silver Edition), one LED display for an instructional video, eight LED displays for video synchronization, and additional lightings. Among 107 cameras, 69 cameras are uniformly placed along the two levels of the dodecagon arc (0.8 m and 1.6 m) for body and cloth, and 38 cameras are place over the frontal hemisphere for face and gaze.

Performance Instructional Video To guide the movements of the participants, we create four instructional videos (∼2.5 minutes). Each video is composed of four sessions. (1) Gaze: a subject is asked to find and look at the requested number tag posted on the camera stage; (2) Face: the subject is asked to follow 20 distinctive dynamic facial expressions (e.g., eye rolling, frowning, and jaw opening); (3) Hand: the subject is asked to follow a series of American sign languages (e.g., counting one to ten, greeting, and daily used words); (4) Body and garment: the subject is asked to follow range of motion, which allows them to move their full body and to follow slow and full speed dance performances curated by a professional choreographer.

Synchronization and Calibration We manually synchronize 107 cameras using LED displays. The maximum synchronization error is up to 15 ms. We use the COLMAP [9] software for camera synchronization, and upgrade the reconstruction to metric scale by using the physical distance between cameras and the ground plane.

B. HUMBI Reconstruction

Given the synchronized multiview image streams, we reconstruct body expressions in 3D.

B.1. 3D Keypoint Reconstruction

Given a set of synchronized and undistorted multiview images, we detect 2D keypoints of face, hand, body (including feet) [1]. Using these keypoints, we triangulate 3D keypoints with RANSAC [2] followed by the non-linear refinement by minimizing reprojection error [3][1]. In the RANSAC process, we apply the length constraint (e.g., symmetry between left and right body) and reason about visibility of keypoints based on confidence of detection, camera proximity, and viewing angle.

B.2. Gaze

We define the moving coordinate of gaze using facial keypoints. Figure 1 illustrates the moving coordinate. The black arrow is gaze direction. The red, green and blue segments are x, y and z-axis of gaze frame. The brown segment is the center axis of the head cylinder. On the right, the orange arrow is the gaze direction. Dark blue box indicates eye region. Blue box wraps face. Yellow area is projection of the cylinder.

B.3. Face

We model $\mathcal{M}_{\text{face}} = f_{\text{face}}(\mathcal{K}_{\text{face}}, \mathcal{I}_{\text{face}})$. We represent a face mesh using Surrey face model [4], which is a 3D

---

*Both authors contributed equally to this work

[1] When multiple persons are detected, we use a geometric verification to identify each subject.

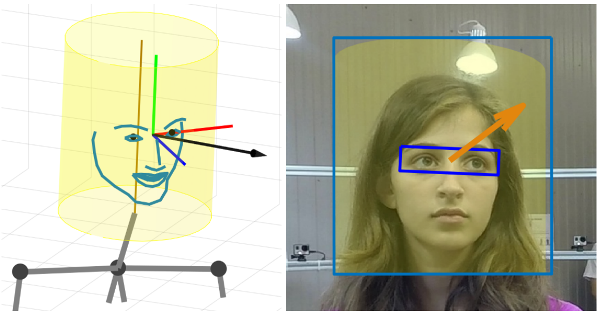

Figure 1: Gaze signals computed by our system (Sec. B.2). (Left) 3D demonstration of captured gaze placed on the black dotted body joints. Black arrow is gaze direction. Red, green and blue segment are x, y and z-axis of gaze frame. Brown segment is the center axis of the head cylinder. (Right) Gaze overlaid on a color image. Orange arrow is gaze direction. Dark blue box indicates eye region. Blue box wraps face. Yellow area is projection of the cylinder.

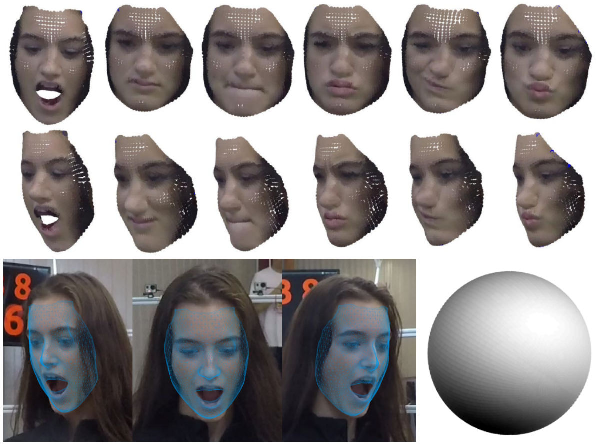

Figure 2: Face reconstruction (Section B.3). (Top) Recovered 3D faces with various expressions (Bottom left) Alignment between projected mesh and subject's face. (Bottom right) Estimated illumination condition.

morphable model (3DMM) defined as:

$$\mathcal{V}_{\text{face}}(\boldsymbol{\alpha}^s,\boldsymbol{\alpha}^e) = \mathbf{S}_0 + \sum_{i=1}^{K_s}\alpha_i^s\mathbf{S}_i + \sum_{i=1}^{K_e}\alpha_i^e\mathbf{E}_i, \tag{1}$$

where $\mathcal{V}_{\text{face}} \in \mathbb{R}^{3D_s}$ is the 3D face vertices, $\mathbf{S}_0$ is the mean-face, $\mathbf{S}_i$ and α_i^s are the i^{th} shape basis and its coefficient, and $\mathbf{E}_i$ and α_i^e are the i^{th} expression basis and its coefficient. D_s is the number of points in the shape model.

The model is fitted to multiview images $\mathcal{I}_{\text{face}}$ by minimizing the following cost:

$$E_{\text{face}} = E^k_{\text{face}} + \lambda^a_{face}E^a_{\text{face}}, \tag{2}$$

where E^k_{face} and E^a_{face} are errors of 3D keypoint and appearance, respectively.

We minimize the geometric error between 3D face model and the reconstructed keypoints:

$$E^k_{\text{face}}(\mathbf{Q},\boldsymbol{\alpha}^s,\boldsymbol{\alpha}^e) = \sum_i^{68}\|\mathcal{K}^i_{\text{face}} - \mathbf{Q}(\overline{\mathbf{V}}^i_{\text{face}})\|^2$$

where $\boldsymbol{\alpha}^s \in \mathbb{R}^{63}$ and $\boldsymbol{\alpha}^e \in \mathbb{R}^6$ are shape and expression coefficients, $\mathcal{K}^i_{\text{face}}$ is i^{th} face keypoint, and $\overline{\mathbf{V}}^i_{\text{face}}$ is the corresponding i^{th} vertex in $\mathbf{V}_{\text{face}}$. $\mathbf{Q}$ is a 6D rigid transformation between the 3DMM in its cannonical coordinate system and the reconstructed model in the world coordinate system.

For appearance fitting, we use text model from Basel Face Model [6]:

$$\mathbf{T} = \mathbf{T}_0 + \sum_{i=1}^{K_t}\alpha_i^t\mathbf{T}_i, \tag{3}$$

where $\mathbf{T} \in \mathbb{R}^{3\times D_s}$ is the 3D face texture, $\mathbf{T}_0$ is the mean texture model, $\mathbf{T}_i$ and α_i^t are the i^{th} texture basis and its coefficient.

The appearance model is combination of texture and illumination: $\mathbf{C} = \mathbf{I}(\mathcal{V}_{\text{face}}, \mathbf{T}, \boldsymbol{\alpha}^h)$ where $\mathbf{C}$ is the RGB color for a 3D face and $\mathbf{I}$ uses Lambertian illumination to estimate the appearance. We model the illumination using the spherical harmonics basis model where $\boldsymbol{\alpha}^h$ is the coefficient for the harmonics. From this, the error of appearance is:

$$E^a_{\text{face}}(\boldsymbol{\alpha}^s,\boldsymbol{\alpha}^e,\boldsymbol{\alpha}^t,\boldsymbol{\alpha}^h) = \sum_j\|\mathbf{c}_j - \phi_j(\mathbf{C})\|^2, \tag{4}$$

where $\phi_j(\mathbf{C})$ is the projection of the appearance $\mathbf{C}$ onto the j^{th} camera, and $\mathbf{c}_j$ is the face appearance in the j^{th} image.

We optimize Equation (2) using a nonlinear least squares solver with ambient light initialization. Figure 2 illustrate the resulting face reconstruction where we compute the shape, expression, texture and reflectance. To learn the consistent shape of the face model for each subject, we infer the maximum likelihood estimate of the shape parameter given the reconstructed keypoints over frames, which allows us to fit to the best model (Figure 2).

B.4. Hand

We model $\mathcal{M}_{\text{hand}}(\boldsymbol{\theta}_h,\boldsymbol{\beta}_h) = f_{\text{hand}}(\mathcal{K}_{\text{face}})$. We represent a hand mesh using the MANO parametric hand model [8], which is composed of 48 pose parameters and 20 shape parameters where $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are the pose and shape parameters, respectively.

We minimize the following objective to model f_{hand}:

$$E_{\text{hand}}(\boldsymbol{\theta},\boldsymbol{\beta}) = E^k_{\text{hand}} + \lambda^\theta_h E^\theta_{\text{hand}} + \lambda^\beta_h E^\beta_{\text{hand}}, \tag{5}$$

where λ_θ and λ_β are weights for pose and shape regularization, respectively.

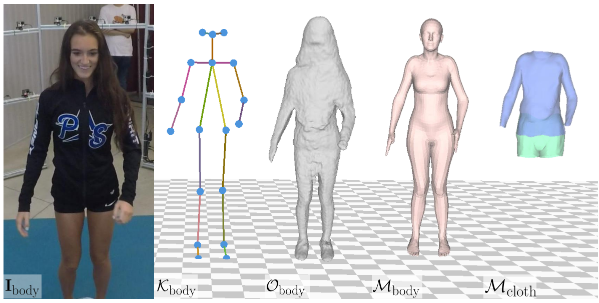
$\mathbf{I}_{\text{body}}$
$\mathcal{K}_{\text{body}}$
$\mathcal{O}_{\text{body}}$
$\mathcal{M}_{\text{body}}$
$\mathcal{M}_{\text{cloth}}$

Figure 3: HUMBI body and cloth reconstruction results.

Given the correspondence between the reconstructed keypoints and the hand mesh, we minimize their error:

$$E^k_{\text{hand}}(\boldsymbol{\theta},\boldsymbol{\beta}) = \sum_i \|\mathcal{K}^i_{\text{hand}} - \mathbf{Q}(\overline{\mathcal{V}}^i_{\text{hand}})\|^2, \tag{6}$$

where $\mathcal{Q}$ is the rigid transformation between the keypoints and the hand mesh model in its canonical coordinate system.

We apply regularization on shape and pose parameters:

$$E^{\theta}_{\text{hand}}(\boldsymbol{\theta},\boldsymbol{\beta}) = \|\boldsymbol{\theta}\|^2, E^{\beta}_{\text{hand}} = \|\boldsymbol{\beta}\|^2. \tag{7}$$

Rigid transformation parameters are firstly estimated by aligning 6 keypoints on palm, then shape and expression parameters are estimated alternatively until converge, followed by nonlinear optimization for all parameters. For the same subject, initially hand mesh of each frame is reconstructed independently. Then shape parameters are fixed as the median values of all frames. Other parameters are optimized, subsequently.

B.5. Body

We model $\mathcal{M}_{\text{body}} = f_{\text{body}}(\mathcal{K}_{\text{body}}, \mathcal{O}_{\text{body}})$. We represent the body expression using a parametric SMPL model [5] and fit to the 3D body keypoints $\mathcal{K}_{\text{body}}$ and the occupancy map $\mathcal{O}_{\text{body}}$ by minimizing the following objective:

$$E_{\text{body}}(\boldsymbol{\alpha}_b,\boldsymbol{\beta}_b,\boldsymbol{\theta}_b) = E^p_{\text{body}} + \lambda^s_b E^s_{\text{body}} + \lambda^r_b E^r_{\text{body}}, \tag{8}$$

where λ^s_b and λ^r_b control the importance of each measurement. $\boldsymbol{\beta}_{\text{b}} \in \mathbb{R}^{10}$ represents the linear shape coefficient, and $\boldsymbol{\alpha}_{\text{b}} \in \mathbb{R}^{72}$ represents Euler angles for the 24 joints (one root joint and 23 relative joints between body parts). $\boldsymbol{\theta}_{\text{body}} \in \mathbb{R}^4$ denotes the translation and scale of the mean body.

We prescribe the correspondence between the pose of SMPL model with 3D body keypoints, i.e., $\overline{\mathcal{V}}^i_{\text{body}}$ is the i^{th} keypoint of the SMPL. E^p_{body} penalizes the distance between the reconstructed 3D body keypoints $\mathcal{K}_{\text{body}}$ and the keypoints of the SMPL $\overline{\mathcal{V}}_{\text{body}}$:

$$E^p_{\text{body}}(\boldsymbol{\alpha}_b,\ \boldsymbol{\theta}_b) = \sum_i \left\|\mathcal{K}^i_{\text{body}} - \overline{\mathcal{V}}^i_{\text{body}}\right\|^2. \tag{9}$$

E^s_{body} encourages the shape of the estimated body model $\mathcal{M}_{\text{body}}$ to be aligned with the outer surface of the occupancy map $\mathcal{O}_{\text{body}}$. We use Chamfer distance to measure the alignment:

$$E^s_{\text{body}}(\boldsymbol{\alpha}_b,\ \boldsymbol{\beta}_b,\ \boldsymbol{\theta}_b) = d_{\text{chamfer}}(\mathcal{O}, \mathcal{V}_{\text{body}}), \tag{10}$$

where d_{chamfer} measures Chamfer distance between two sets of point clouds.

E^r_{body} penalizes the difference between the estimated shape $\boldsymbol{\beta}_b$ and the subject-aware mean shape $\boldsymbol{\beta}^{\text{prior}}_b$ as follows:

$$E^r_{\text{body}}(\boldsymbol{\beta}_b;\boldsymbol{\beta}^{\text{prior}}_b) = \left\|\boldsymbol{\beta}_b - \boldsymbol{\beta}^{\text{prior}}_b\right\|^2. \tag{11}$$

This prevents unrealistic shape fitting due to the estimation noise/error, e.g., long hair covering body. To obtain the shape prior $\boldsymbol{\beta}^{prior}_b$, we solve the Eq. (8) without E^{body}_r using the recovered volumes of the same subject and take the median $\boldsymbol{\beta}_b$ for robustness.

B.6. Garment

We model a garment fitting function $\mathcal{M}_{\text{cloth}} = f_{\text{cloth}}(\mathcal{M}_{\text{body}}, \mathcal{O}_{\text{body}})$ by representing the garment with an in-house mesh model $\mathcal{M}_{\text{cloth}}$. The assumption of the minimally clothed body shape [7] is made. We minimize the following objective:

$$E_{\text{cloth}}(\mathbf{R}_c, \mathbf{t}_c) = E^b_{\text{cloth}} + \lambda^o_c E^o_{\text{cloth}} + \lambda^r_c E^r_{\text{cloth}}, \tag{12}$$

where λ^o_c and λ^r_c control the importance of each measurement.

We manually establish the set of correspondences between $\mathcal{M}_{\text{body}}$ and $\mathcal{M}_{\text{cloth}}$ that move approximately the same way. E^b_{cloth} measures the correspondence error:

$$E^b_{\text{cloth}}(\mathcal{V}_{\text{cloth}}) = \sum_i \|\overline{\mathcal{V}}^i_{\text{body}} - \overline{\mathcal{V}}^i_{\text{cloth}}\|^2, \tag{13}$$

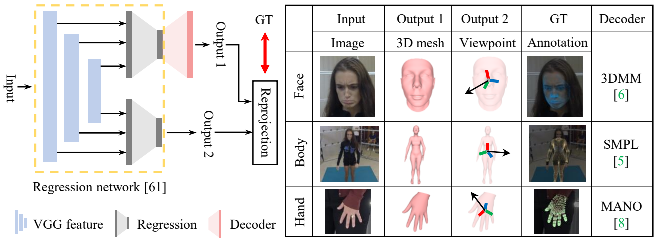

Figure 4: The training setup for 3D mesh prediction from a single image.

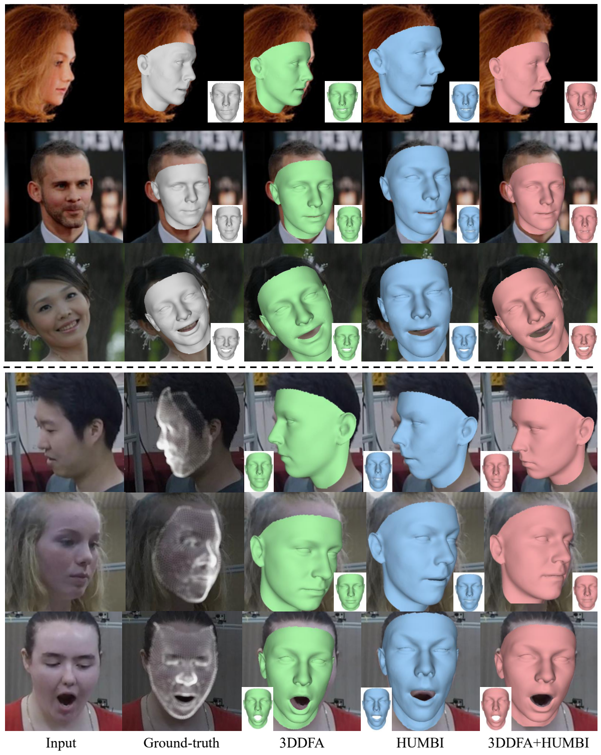

Figure 5: The qualitative results of the monocular 3D face prediction network trained with different dataset combination. The top and bottom show the testing on the external and HUMBI Face respectively.

where $\overline{\mathcal{V}}_{\text{body}}$ and $\overline{\mathcal{V}}_{\text{cloth}}$ are the corresponding vertices.

E^o_{cloth} measures the Chamfer distance to align $\mathcal{M}_{\text{cloth}}$ with $\mathcal{O}_{\text{body}}$:

$$E^o_{\text{cloth}}(\mathcal{V}_{\text{cloth}}) = d_{\text{chamfer}}(\mathcal{O}_{\text{body}}, \mathcal{V}_{\text{cloth}}). \tag{14}$$

E^r_{cloth} is the spatial regularization (Laplacian) that prevents from reconstructing unrealistic cloth structure by penalizing a non-smooth and non-rigid vertex with respect to its neighboring vertices [10]:

$$E^r_{\text{cloth}} = \nabla^2 \mathcal{M}_{\text{cloth}}. \tag{15}$$

C. Training Mesh Prediction Network

To train the mesh prediction function of each body expression (i.e., face, hand, and body described in Section 4.1-4.3 of the main paper), we use the recent neural network [11] that can regress a single image to the body model parameters, e.g., SMPL body shape and pose coefficients, and camera viewpoint. In Figure 4, the encoder is implemented with [11], and the decoder with the pre-trained weights of each body model, i.e., 3DMM [6] for face, SMPL [5] for body, and MANO [8] for hand. The network is trained by minimizing the reprojection error where only the regression network is newly trained. The training details are described in Figure 4.

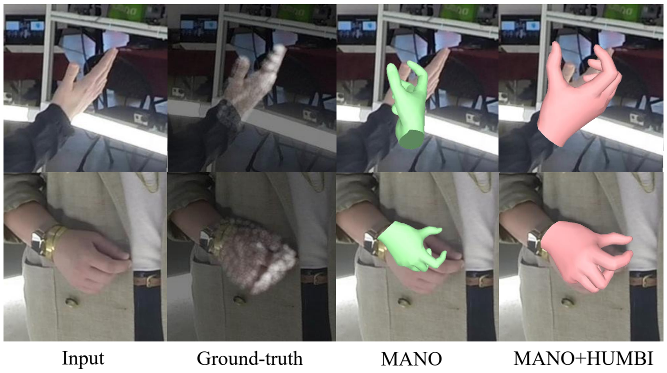

Figure 6: Monocular 3D hand mesh prediction results tested on HUMBI Hand.

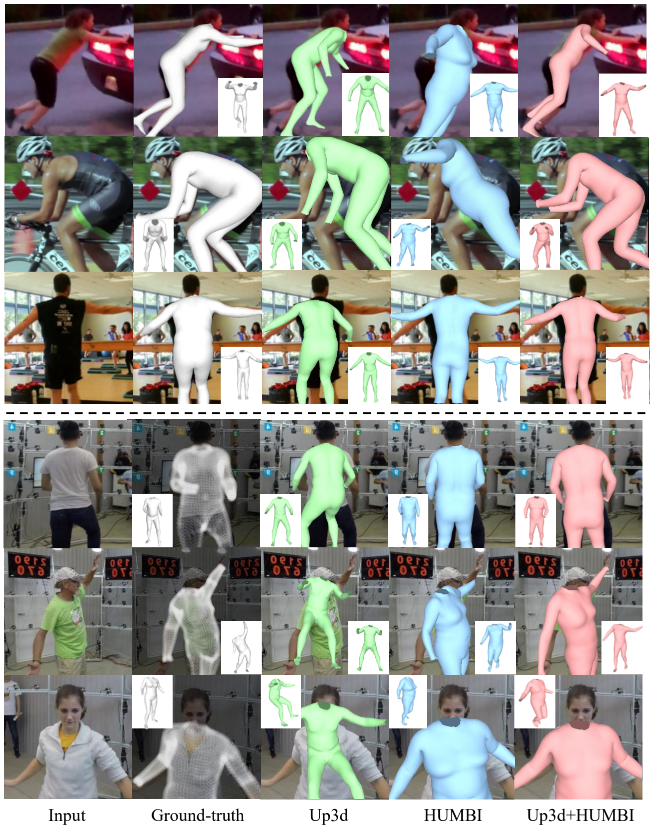

Figure 7: The qualitative results of the monocular 3D body prediction network trained with different dataset combination. The top and bottom show the results tested on UP-3D and HUMBI Body respectively.

D. More Results

D.1. Mesh Prediction Results

We use a recent CNN model to evaluate HUMBI as introduced in Section C. The qualitative evaluation on single view prediction is shown in Figure 5 (face), Figure 6 (hand), and Figure 7 (body).

D.2. Garment Reconstruction Accuracy

We provide additional evaluation of view-dependent garment silhouette accuracy measured by the Chamfer distance between the annotated and the reprojected garment boundary in 2D. We pick a half-sleeve shirts and half pants models as a representative garment of top and bottom and measure the accuracy from each camera view that has different angle with respect to the most frontal camera. On average in Figure 8, the silhouette error seen from the side view (11 pixels) is higher than the frontal (7.5 pixels) and rear views (8 pixels).

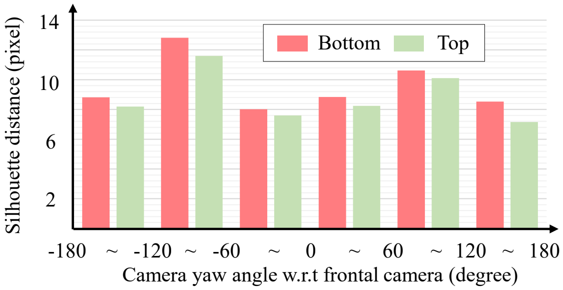

Figure 8: Garment silhouette error.

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 1

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *ACM Comm.*, 1981. 1

[3] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 1

[4] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework. In *VISIGRAPP*, 2016. 1

[5] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *TOG*, 2015. 3, 4

[6] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. *AVSS*, 2009. 2, 4

[7] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *TOG*, 2017. 3

[8] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH*, 2017. 2, 4

[9] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[10] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, 2007. 4

[11] J. S. Yoon, T. Shiratori, S.-I. Yu, and H. S. Park. Self-supervised adaptation of high-fidelity face models for monocular performance tracking. In *CVPR*, 2019. 4