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

Figure 1: We present a new large dataset of multiview human body expressions for modeling view-specific appearance and geometry. 107 synchronized cameras capture the expressions of 772 distinctive subjects. We focus on five elementary expressions: face (blue), gaze (yellow), hand (pink and purple), body (light orange), and garment including top (light blue) and bottom (light green).

Abstract

This paper presents a new large multiview dataset called HUMBI for human body expressions with natural clothing. The goal of HUMBI is to facilitate modeling view-specific appearance and geometry of gaze, face, hand, body, and garment from assorted people. 107 synchronized HD cameras are used to capture 772 distinctive subjects across gender, ethnicity, age, and physical condition. With the multiview image streams, we reconstruct high fidelity body expressions using 3D mesh models, which allows representing view-specific appearance using their canonical atlas. We demonstrate that HUMBI is highly effective in learning and reconstructing a complete human model and is complementary to the existing datasets of human body expressions with limited views and subjects such as MPII-Gaze, Multi-PIE, Human3.6M, and Panoptic Studio datasets.

1. Introduction

We express sincere intent, emotion, and attention through our *honest body signals* [50], including gaze, facial expression, and gestures. Modeling and photorealistic rendering of such body signals are, therefore, the core enabler of authentic telepresence. However, it is challenging due to the complex physical interactions between texture, geometry, illumination, and viewpoint (e.g., translucent skins, tiny wrinkles, and reflective fabric). Recently, pose- and view-specific models by making use of a copious capacity of neural encoding [6,39] substantially extend the expressibility of existing linear models [16]. So far, these models have been constructed by a sequence of the detailed scans of a target subject using dedicated camera infrastructure (e.g., multi-camera systems [7, 25, 73]). Looking ahead, we would expect a new versatile model that is applicable to the general appearance of assorted people without requiring the massive scans for every target subject.

Among many factors, what are the core resources to build such a generalizable model? We argue that the data that can span an extensive range of appearances from numerous shapes and identities are prerequisites. To validate our conjecture, we present a new dataset of human body expressions called *HUMBI* (HUman Multiview Behavioral Imaging) that pushes to two extremes: views and subjects. As of Nov 2019[1], the dataset is composed of 772 distinctive subjects with natural clothing across diverse age, gender, ethnicity, and physical condition captured by 107 HD synchronized cameras (68 cameras facing at frontal body). Comparing to existing datasets for human body expressions such as CMU Panoptic Studio [24, 26], MPII [51, 52], and

*Both authors contributed equally to this work

[1]In a contract with public event venues, the dataset is expected to grow every year.

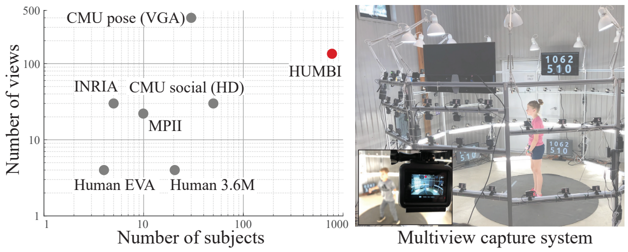


Figure 2: We present HUMBI that pushes towards two extremes: views and subjects. The view-specific appearance measured by 107 HD cameras regarding five elementary body expressions for 772 distinctive subjects.

INRIA [31], HUMBI presents the unprecedented scale visual data (Figure 2) that are ideal for learning the detailed appearance and geometry of five elementary human body expressions: gaze, face, hand, body, and garment (Figure 1).

Our analysis shows that HUMBI is effective. We make use of vanilla convolutional neural networks (CNN) to learn view-invariant 3D pose from HUMBI, which quantitatively outperforms the counterpart models trained by existing datasets with limited views and subjects. More importantly, we show that HUMBI is *complementary* to such datasets, i.e., the trained models can be substantially improved by combining with these datasets.

The main properties of HUMBI are summarized below. (1) Complete: it captures the total body, including gaze, face, hand, foot, body, and garment to represent holistic body signals [28], e.g., perceptual asynchrony between the face and hand movements. (2) Dense: 107 HD cameras create a dense light field that observe the minute body expressions with minimal self-occlusion. This dense light field allows us to model precise appearance as a function of view [39]. (3) Natural: the subjects are all voluntary participants (no actor/actress/student/researcher). Their activities are loosely guided by performance instructions, which generates natural body expressions. (4) Diverse: 772 distinctive subjects with diverse clothing styles, skin colors, time-varying geometry of gaze/face/body/hand, and range of motion. (5) Fine: with multiview HD cameras, we reconstruct the high fidelity 3D model using 3D meshes, which allows representing view-specific appearance in its canonical atlas.

2. Related Work

We briefly review the existing datasets for modeling human body expressions: gaze, face, hand, body, and garment. These datasets are summarized in Table 1.

Gaze Columbia Gaze dataset [62] and UT-Multiview dataset [64] have been captured in a controlled environments where the head poses are fixed. In subsequent work, such constraints have been relaxed. Eyediap dataset [43] captured gaze while allowing head motion, providing natural gaze movements. MPII-Gaze dataset [81] measured in-the-wild gaze from laptops, including 214K images across 15 subjects. This contains a variety of appearance and illumination. RT-GENE dataset [18] takes a step further by measuring free-ranging point of regard where the ground truth was obtained by using motion capture of mobile eye-tracking glasses.

Face 3D Morphable Model (3DMM) [10] was constructed by 3D scans of large population to model the complex geometry and appearance of human faces. For instance, 3D faces were reconstructed by leveraging facial landmarks [8, 29, 36, 55, 57], and dense face mesh [17, 68]. Notably, 3DMM is fitted to 60K samples from several face alignment datasets [8, 42, 56, 82, 85] to create the 300W-LP dataset [84]. For facial appearance, a deep appearance model [39] introduces view-dependent appearance using a conditional variational autoencoder, which outperforms linear active appearance model [16].

Hand Dexterous hand manipulation frequently introduces self-occlusion, which makes building a 3D hand pose dataset challenging. A depth image that provides trivial hand segmentation in conjunction with tracking has been used to establish the ground truth hand pose [65–67, 69]. However, such approaches still require intense manual adjustments. This challenge was addressed by making use of graphically generated hands [44, 45, 86], which may introduce a domain gap between real and synthetic data. For real data, an auxiliary input such as magnetic sensors was used to precisely measure the joint angle and recover 3D hand pose using forward kinematics [74, 78]. Notably, a multi-camera system has been used to annotate hands using 3D bootstrapping [61], which provided the hand annotations for RGB data. FreiHAND [15] leveraged MANO [53] mesh model to represent dense hand pose.

Body Markerless motion capture is a viable solution to measure dense human body expression at high resolution. For example, multi-camera systems have been used to capture a diverse set of body poses, e.g., actors and actresses perform a few scripted activities such as drinking, answering cellphone, and sitting [23, 60]. Natural 3D human behaviors were captured in the midst of the role-playing of social events from a multiview system [27], while those events inherently involve with a significant occlusion by people or objects that inhibit modeling a complete human body. Further, a 4D scanner [11, 52] enabled high resolution body capture to construct a parametric human models, e.g., SMPL [40]. Notably, image-to-surface correspondences on 50K COCO images [38] enabled modeling humans from a single view image [32]. Further, rendering of human model in images could alleviate annotation efforts [70].

Clothes Previous works have proposed to capture the natural cloth deformation in response to human body movement. Cloth regions were segmented in 3D using multiview reconstruction [13, 75]. To ensure the same topology when segmenting the cloth from 3D reconstruction, the SMPL body model can be used to parametrize cloth motion, which produces physically plausible cloth geometry while preserving

Dataset	# of subjects	Measurement method	Gaze	Face	Hand	Body	Cloth
Columbia Gaze [62]	56	5 cameras	✓(fixed)				
UT-Multiview [64]	50	8 cameras	✓(fixed)				
Eyediap [43]	16	1 depth camera and 1 HD camera	✓(free)				
MPII-Gaze [81]	15	1 camera	✓(free)				
RT-GENE [18]	17	eyetracking device	✓(free)				
CMU Multi-PIE [19]	337	15 cameras		✓			
3DMM [10]	200	3D scanner		✓			
BFM [49]	200	3D scanner		✓			
ICL [12]	10,000	3D scanner		✓			
NYU Hand [69]	2 (81K samples)	Depth camera			✓		
HandNet [74]	10 (213K samples)	Depth camera and magnetic sensor			✓		
BigHand 2.2M [78]	10 (2.2M samples)	Depth camera and magnetic sensor			✓		
RHD [86]	20 (44K samples)	N/A (synthesized)			✓		
STB [80]	1 (18K samples)	1 pair of stereo cameras			✓		
FreiHand [15]	N/A (33K samples)	8 cameras			✓		
CMU Mocap	$\sim$100	Marker-based				✓	
CMU Skin Mocap [47]	$<$10	Marker-based		✓		✓	
INRIA [31]	N/A	Markerless (34 cameras)				✓	✓(natural)
Human EVA [60]	4	Marker-based and Markerless (4-7 cameras)				✓	
Human 3.6M [23]	11	Markerless (depth camera and 4 HD cameras)				✓	
Panoptic Studio [27,61]	$\sim$100	Markerless (31 HD and 480 VGA cameras)			✓	✓	
Dyna [52]	10	Markerless (22 pairs of stereo cameras)				✓	
ClothCap [51]	10	Markerless (22 pairs of stereo cameras)					✓(synthesized)
BUFF [79]	5	Markerless (22 pairs of stereo cameras)				✓	✓(natural)
3DPW [71]	7	Marker-based (17 IMUs) and Markerless (1 camera + 3D scanner)				✓	✓(natural)
TNT15 [72]	4	Marker-based (10 IMUs) and Markerless (8 cameras + 3D scanner)				✓	
D-FAUST [11]	10	Markerless (22 pairs of stereo cameras)				✓	
HUMBI	772	Markerless (107 HD cameras)	✓(free)	✓	✓	✓	✓(natural)

Table 1: Human body expression datasets.

wrinkle level details [51].

Our Approach Unlike existing datasets focusing on each body expressions, HUMBI is designed to span geometry and appearance of total body expressions from a number of distinctive subjects using a dense camera array. Our tera-scale multiview visual data provide a new opportunity to generalize pose- and view-specific appearance.

3. HUMBI

HUMBI is composed of 772 distinctive subjects captured by 107 synchronized HD cameras. 69 cameras are uniformly distributed across dodecagon frame with 2.5m diameter along the two levels of an arc (0.8 m and 1.6 m) where the baseline between adjacent cameras is approximately 10° (22 cm). Another 38 cameras are distributed across the frontal quadrant of the dodecagon frame (average baseline: 10 cm) to densify cameras used for capturing face/gaze. The dataset includes the five elementary body expressions: gaze, face, hand, body, and garment. We use COLMAP [59] to calibrate cameras, and upgrade the reconstruction to the metric scale using physical camera baselines. Notable subject statistics includes: evenly distributed gender (50.7% female; 49.3% male); a wide range of age groups (11% of thirties, 29% of twenties, and 26% of teenagers); diverse skin colors (black, dark brown, light brown, and white); various styles of clothing (dress, short-/long-sleeve t-shirt, jacket, hat, and short-/long-pants). The statistics are summarized in Figure 3. In this section, we focus on the resulting computational representations while deferring the detailed description of reconstruction approaches to Appendix.

Notation We denote our representation of human body expressions as follows:

- Images: $\mathcal{I} = \{\mathbf{I}_i\}$ is a set of multiview images.
- 3D keypoints: $\mathcal{K}$.
- 3D mesh: $\mathcal{M} = \{\mathcal{V}, \mathcal{E}\}$.
- 3D occupancy map: $\mathcal{O} : \mathbb{R}^3 \rightarrow \{0, 1\}$ that takes as input 3D voxel coordinate and outputs binary occupancy.
- Appearance map: $\mathcal{A} : \mathbb{R}^2 \rightarrow [0, 1]^3$ that takes as input atlas coordinate (UV) and outputs normalized RGB values.

Keypoint 3D keypoints on face ($\mathcal{K}_{\rm face}$), hands ($\mathcal{K}_{\rm hand}$), and body including feet ($\mathcal{K}_{\rm body}$) are reconstructed by triangulating 2D human keypoint detections [14] with RANSAC, followed by a nonlinear refinement minimizing geometric reprojection error. When multiple humans are visible, we localize each subject via geometric verification.

3.1. Gaze

HUMBI Gaze contains $\sim$93K images (4 gaze directions $\times$ $\sim$30 views per subject). We represent gaze geometry using a unit 3D vector $\mathbf{g} \in \mathbb{S}^2$ with respect to the moving head coordinate system.

The head coordinate is defined as follows. The origin is the center of eyes, $\mathbf{o} = (\mathbf{p}_l + \mathbf{p}_r)/2$ where $\mathbf{p}_l, \mathbf{p}_r \in \mathbb{R}^3$ are left and right eye centers. The x-axis is the direction along the line joining the two eye centers, $(\mathbf{p}_l - \mathbf{o})/\|\mathbf{p}_l - \mathbf{o}\|$; the z-axis is the direction perpendicular to the plane made of $\mathbf{p}_l$, $\mathbf{p}_r$, and $\mathbf{p}_m$ where $\mathbf{p}_m$ is the center of the mouth, orienting towards the hind face; y-axis is defined as a vector orthogonal to both x- and z-axes under right-hand rule constraint.

For eye appearance, we provide two representations: (1) normalized eye patches and (2) pose-independent appearance map. For the normalized eye patches, we warp an eye patch region such that the orientation and distance remain constant across views. RGB values are histogram-equalized. For appearance, we select vertices of eye region in the Surrey face model [22] to build a canonical atlas coordinate (UV) for each eye. We represent view-specific

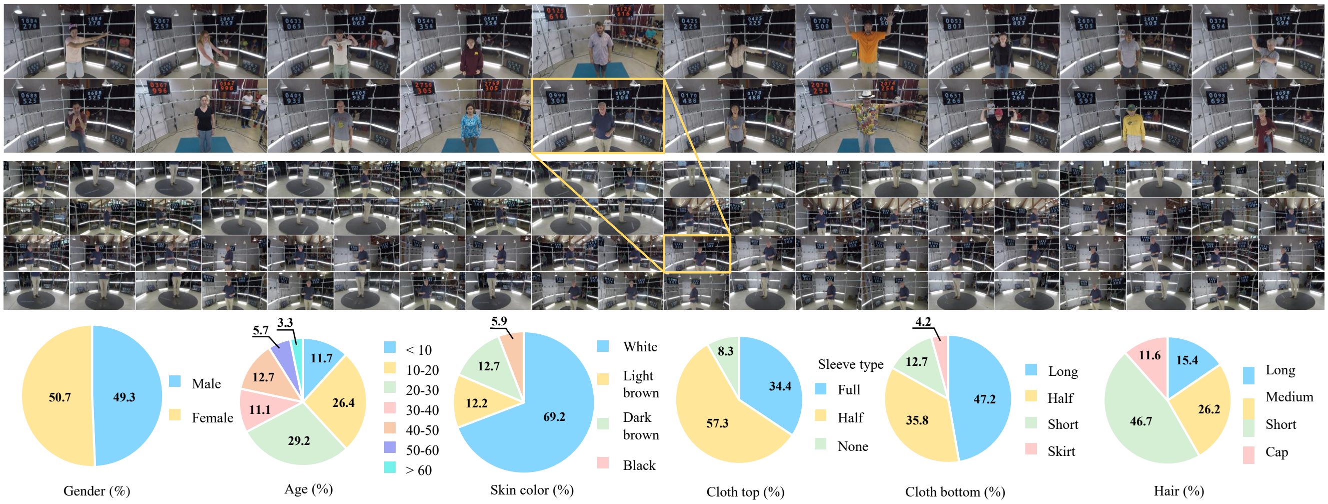


Figure 3: (Top and bottom) HUMBI includes 772 distinctive subjects across gender, ethnicity, age, clothing style, and physical condition, which generates diverse appearance of human expressions. (Middle) For each subject, 107 HD cameras capture her/his expressions including gaze, face, hand, body, and garment.

appearance map $\mathcal{A}_{\rm gaze}$ by projecting pixels in the image onto that the atlas coordinate. Figure 4(a) illustrates view-specific appearance across views with median and variance of appearance. The variance map shows that the appearance is highly dependent on viewpoint in particular in the iris region.

3.2. Face

HUMBI Face contains $\sim$17.3M images (330 frames $\times$ 68 views per subject). We represent face geometry using a 3D blend shape model $\mathcal{M}_{face}$ (Surrey [22]) with 3,448 vertices and 6,736 faces. We reconstruct the shape model using 68 facial keypoints ($\mathcal{K}_{\rm face}$) and the associated multiview images ($\mathcal{I}_{\rm face}$), i.e., $\mathcal{M}_{\rm face} = f_{\rm face}(\mathcal{K}_{\rm face}, \mathcal{I}_{\rm face})$ where $f_{\rm face}$ is a face alignment function. We align the face model by minimizing reprojection error over shape, expression, illumination, and texture parameters (see Appendix). Given the reconstructed face mesh model, we construct a view-specific appearance map $\mathcal{A}_{\rm face}$ by projecting pixels in the image onto its canonical atlas coordinate. For each view, the projection map between the image and atlas coordinate is established through the corresponding 3D locations in the reconstructed mesh with bilinear interpolation. Figure 4(b) illustrates view-specific appearance across views with median and variance of appearance. The variance map shows that the appearance is dependent on views, e.g. the regions of salient landmarks such as eye, eyebrows, nose, and mouth, which justifies the necessity of view-specific appearance modeling [39].

3.3. Hand

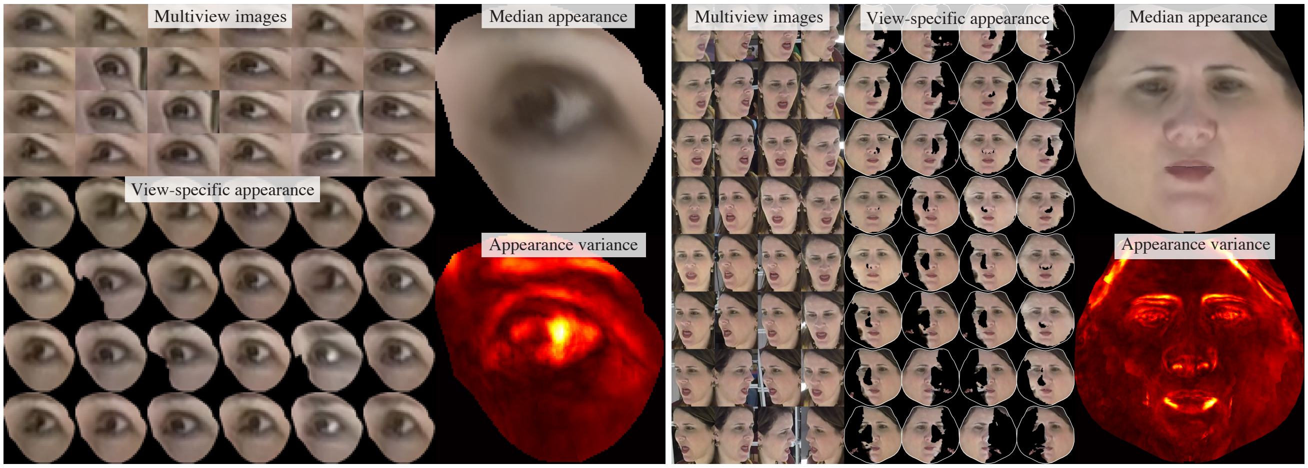
HUMBI Hand contains $\sim$24M images (290 frames $\times$ 68 views per subject). We represent hand geometry using a 3D parametric model $\mathcal{M}_{\rm hand}$ (MANO [53]) with 778 vertices and 1,538 faces. We reconstruct the mesh model using hand keypoints ($\mathcal{K}_{\rm hand}$ with 21 keypoints), i.e., $\mathcal{M}_{\rm hand} = f_{\rm hand}(\mathcal{K}_{\rm face})$, where $f_{\rm hand}$ is a hand alignment function. We align the hand model to multiview images by minimizing the Euclidean distance between hand keypoints and the corresponding pose of the mesh model with a L_2 parameter regularization. To learn the consistent shape of the hand model for each subject, we infer the maximum likelihood estimate of the shape parameter given the reconstructed keypoints over frames (see Appendix). Given the reconstructed hand mesh model, we construct a view-specific appearance map $\mathcal{A}_{\rm hand}$ by projecting pixels in an image onto the canonical atlas coordinate. Figure 4(c) illustrates view-specific appearance across views with median and variance of appearance. The variance map shows that the appearance is dependent on view points.

3.4. Body

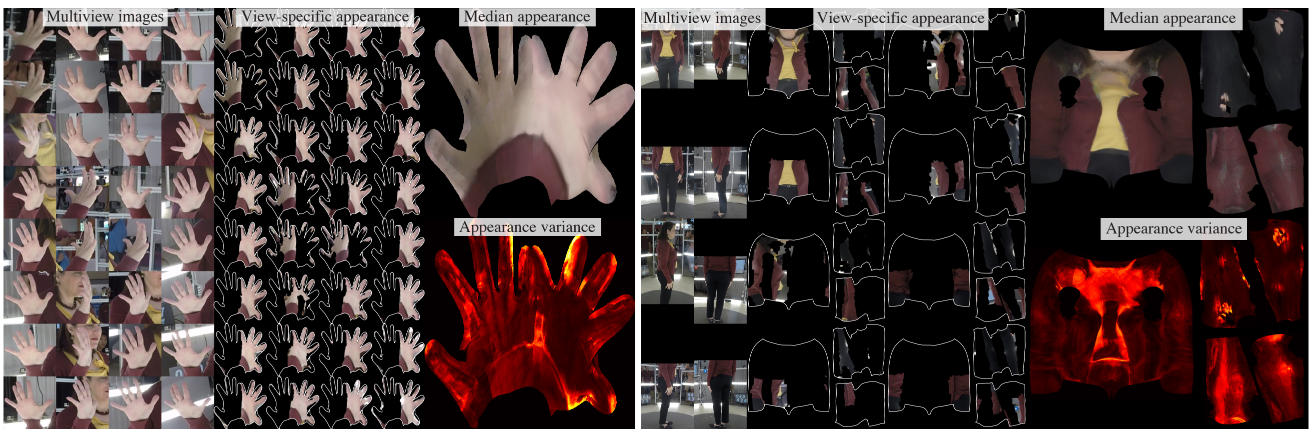
Each subject performs a sequence of motion and dance performance, which constitutes $\sim$26M images. Given a set of multiview images at each time instant, we reconstruct a mesh model $\mathcal{M}_{\rm body}$ using body keypoints $\mathcal{K}_{\rm body}$, and occupancy map $\mathcal{O}_{\rm body}$, i.e., $\mathcal{M}_{\rm body} = f_{\rm body}(\mathcal{K}_{\rm body}, \mathcal{O}_{\rm body})$ where $f_{\rm body}$ is a alignment function that matches the surface of $\mathcal{M}_{\rm body}$ to the outer surface of the occupancy map while minimizing the distance between the reconstructed keypoints $\mathcal{K}_{\rm body}$ and the underlying pose of the mesh (see Appendix). We use the SMPL parametric model [40] that is composed of 4,129 vertices and 7,999 faces without hand and head vertices.

Shape-from-silhouette[2] [34] is used to reconstruct the occupancy map $\mathcal{O}_{\rm body}$. The occupancy map is generated by human body segmentation [37]. As a by-product, the semantics (i.e., head, torso, upper arm, lower arm, upper leg,

[2]MultiView stereo [59] is complementary to the occupancy map.



(a) Gaze appearance (b) Face appearance



(c) Hand appearance (d) Body appearance

Figure 4: View-specific appearance rendered from multiview images with median appearance and variance for (a) gaze, (b) face, (c) hand, (d) body.

and lower leg) can be labeled at each location in the occupancy map by associating with the projected body label [76] as shown in Figure 5.

Given the reconstructed body mesh model, we construct a view-specific appearance map $\mathcal{A}_{\text{body}}$ by projecting pixels in an image onto the canonical atlas coordinate. Figure 4(d) illustrates view-specific appearance across views with median and variance of appearance. The variance map shows that the appearance is dependent on view points.

3.5. Garment

Given the body reconstruction, we represent the garment geometry using a garment mesh model $\mathcal{M}_{\text{cloth}}$ as similar to [51]. An alignment function $\mathcal{M}_{\text{cloth}} = f_{\text{cloth}}(\mathcal{M}_{\text{body}}, \mathcal{O}_{\text{body}})$ is used to reconstruct the cloth mesh model from the body model and occupancy map. A set of fiducial correspondences between the cloth and body meshes are predefined, which are used as control points for cloth deformation. The deformed cloth is matched to the outer surface of the occupancy map with a Laplacian regularization [63] (see Appendix). Three garment topologies for each cloth piece are used, i.e., tops: sleeveless shirts (3,763 vertices and 7,261 faces), T-shirts (6,533 vertices, 13,074 faces), and long-sleeve shirts (8,269 vertices and 16,374 faces), and bottoms: short (3,975 vertices and 7,842 faces), medium (5,872 vertices and 11,618 faces), and long pants (11,238 vertices and 22,342 meshes), which are manually matched to each subject.

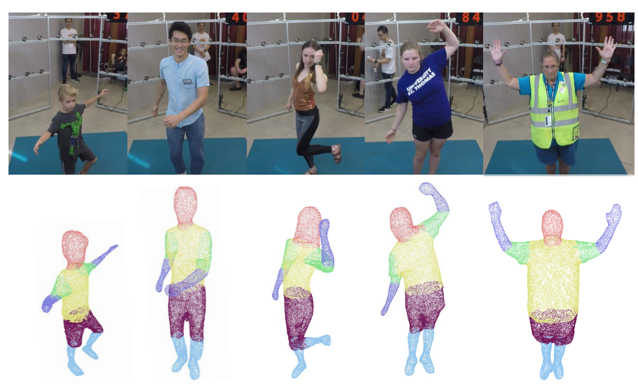


Figure 5: We reconstruct the body occupancy map and its outer surface using shape-from-silhouette and associate the point cloud with body semantics (head, body, arms, and legs).

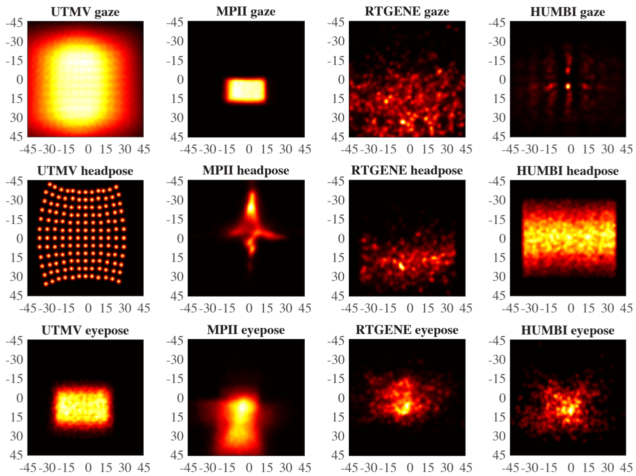


Figure 6: Distribution of head pose, gaze and eye pose in normalized space for MPII-Gaze, UT-Multiview, RT-GENE and HUMBI. Horizontal and vertical axis represent yaw and pitch angle respectively (unit: degree).

Bias / Variance ↘	UTMV	MPII	RTGENE	HUMBI
Gaze	**7.43** / **33.09**	8.80 / 10.10	19.35 / <u>31.71</u>	<u>7.70</u> / 30.01
Headpose	<u>4.20</u> / **29.28**	12.51 / 16.04	17.97 / 22.48	**1.42** / <u>24.77</u>
Eyepose	<u>8.43</u> / 15.40	20.81 / <u>19.02</u>	**3.21** / 17.49	8.78 / **19.04**
Average	<u>6.69</u> / **25.93**	14.04 / 15.05	13.51 / 23.90	**5.98** / <u>24.61</u>

Table 2: Bias and variance analysis of the distribution of head pose, gaze and eye pose (unit: degree, smallest bias and largest variance in bold, second with underline).

4. Evaluation

We evaluate HUMBI in terms of generalizability, diversity, and accuracy. For the generalizability, we conduct the cross-data evaluation on tasks of single view human reconstruction, e.g., monocular 3D face mesh prediction. For diversity, we visualize the distribution of HUMBI, e.g., gaze direction distribution along the yaw and pitch angle. For the accuracy, we measure how the number of cameras affects the quality of reconstruction. More evaluations can be found in Appendix.

4.1. Gaze

Benchmark Datasets We use three benchmark datasets: (1) MPII-Gaze (MPII) [81] contains 213,659 images from 15 subjects, which was captured under the scenarios of everyday laptop use. (2) UT-Multiview (UTMV) [64] is composed of 50 subjects with 160 gaze directions captured by 8 monitor-mounted cameras. Using the real data, the synthesized images from 144 virtual cameras are augmented. (3) RT-GENE [18] contains 122,531 images of 15 subjects captured by eye-tracking glasses.

Distribution of Gaze Directions To characterize HUMBI Gaze, we visualize three measures in Figure 6: (1) gaze pose: the gaze direction with respect to camera pose; (2) head pose: the head orientation with respect to the camera pose; and (3) eye pose: the gaze direction with respect to the head. HUMBI covers a wide and continuous range of head poses, due to numerous views and natural head movements by many subjects. The yaw and pitch of gaze and eye poses are distributed uniformly across all angles. The quantitative analysis of the bias and variance of the gaze distribution is summarized in Table 2. HUMBI shows the smallest average bias (5.98° compared to 6.69°-14.04° from other datasets) and second-largest average variance (24.61° compared to 25.93° of UTMV). Notice that UTMV is a synthesized dataset while HUMBI is real.

Testing \ Training	MPII	UTMV	HUMBI	MPII + HUMBI	UTMV + HUMBI
MPII	6.1±3.3	11.8±6.6	8.8±4.8	7.4±4.1	7.7±4.6
UTMV	23.3±9.4	5.0±3.2	8.2±4.5	9.4±5.1	5.4±3.2
HUMBI	23.7±13.7	14.6±10.3	7.9±5.4	8.9±6.2	8.0±5.4

Table 3: The mean error of 3D gaze prediction for the cross-data evaluation (unit: degree).

Monocular 3D Gaze Prediction To validate the generalizability of HUMBI Gaze, we use an existing gaze detection network [81] to conduct a cross-data evaluation. We randomly choose ∼25K images (equally distributed among subjects) as experiment set for each dataset. One dataset is used for training and others are used for testing. Each data sample is defined as $\{(\mathbf{e}_c, \mathbf{h}_c), \mathbf{g}_c\}$, where $\mathbf{e}_c \in \mathbb{R}^{36\times 60}, \mathbf{h}_c \in \mathbb{R}^2, \mathbf{g}_c \in \mathbb{R}^2$ are normalized eye patch, yaw and pitch angle of head pose, and gaze direction with respect to a virtual camera c. The detection network is trained to minimize the mean squared error of gaze yaw and pitch angles. We conduct a self-data evaluation for each dataset with 90%/10% of training/testing split. Table 3 summarize the experiment results. The detector trained by MPII and UTMV shows weak performance on cross-data evaluation comparing to HUMBI with 3°-16° margin. HUMBI exhibits strong performance on cross-data evaluation with minimal degradation (less than 1° drop). Also, UTMV + HUMBI and MPII + HUMBI outperform each alone by a margin of 4.1° and 13.9° when tested on the third dataset MPII and UTMV respectively, showing that HUMBI is complementary to UTMV and MPII.

4.2. Face

Benchmark Dataset We use 3DDFA [84] that provides ∼6K 2D-3D pairs of the 3D face geometry and the associated images. We use 90%/10% of training/testing split. The base face model of 3DDFA is the Basel model [49], which is different from our face model (Surrey [22]). We manually pre-define the correspondences between two models in the canonical coordinates.

Monocular 3D Face Mesh Prediction We evaluate HUMBI Face by predicting a 3D face mesh using a recent mesh reconstruction network [77]. The network encoder directly regresses the 3D face shape and head pose from a single view image. We modify the decoder to accommodate the differentiable Basel model. We train the network

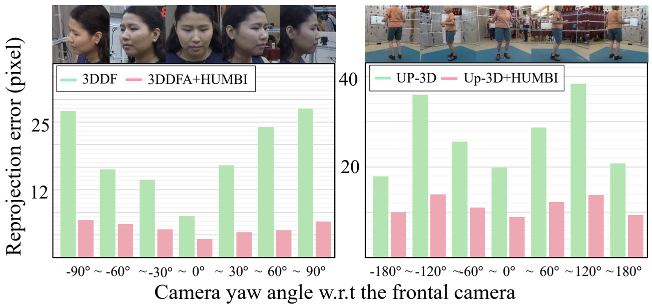


Figure 7: We measure viewpoint dependency of a face/body mesh reconstruction model trained by multiple datasets. Augmenting HUMBI substantially reduce the view dependency.

Testing \ Training	3DDFA	HUMBI	3DDFA+HUMBI
3DDFA	7.1±6.4	20.7±7.1	**4.3±6.6**
HUMBI	23.5±13.9	13.3±13.7	**8.4±12.2**

Table 4: The mean error of 3D face mesh prediction for cross-data evaluation (unit: pixel).

with three dataset combinations, i.e., 3DDFA, HUMBI, and 3DDFA+HUMBI, and for each training, we minimize the loss of the reprojection error with weak perspective projection model. To measure the accuracy, we use the reprojection error scaled to the input image resolution (256 pixel). Table 4 summarize the results. From the results of 3DDFA+HUMBI, the prediction accuracy is improved from both datasets (2.8 pixels from 3DDFA and 4.9 pixels from HUMBI) by combining two datasets, which indicates the complementary nature of HUMBI. Due to the multiview images in HUMBI, the network can learn the view-invariant geometric representation, which allows precise reconstruction even with considerable occlusion as shown in Figure 7.

4.3. Hand

Benchmark Datasets We use three benchmark datasets: (1) Rendered Handpose Dataset (RHD) [86] is a synthesized hand dataset containing 44K images built from 20 freely available 3D models performing 39 actions. (2) Stereo Hand Pose Tracking Benchmark (SHPTB) [80] is a real hand dataset captured by a stereo rgb camera rig. (3) FreiHAND [15] is a multi-view real hand dataset captured by 8 cameras. (4) ObMan [21] is a large scale synthetic hand mesh dataset with associated 2D images (141K pairs). We use previous two datasets for the hand keypoint evaluation and the last one for the hand mesh evaluation.

Monocular 3D Hand Pose Prediction To validate HUMBI Hand, we conduct a cross-data evaluation for the task of the 3D hand pose estimation from a single view image, where we use a recent hand pose detector [86]. We train and evaluate the model trained by each dataset and a combination of HUMBI and each other dataset. The results are summarized in Table 5. We use area under PCK curve (AUC) in an error range of 0-20mm as the metric. It show that HUMBI is more generalizable for predicting 3D hand pose than other three dataset (by a margin of 0.02-0.16 AUC). Moreover, HUMBI is complementary to other datasets and the performance of model trained by another dataset alone is increased with HUMBI (by a margin of 0.04-0.12 AUC).

Testing \ Training	S	R	F	H	S+H	R+H	F+H
STB (S)	0.72	0.40	0.22	0.47	0.40	0.52	0.44
RHD (R)	0.16	0.59	0.26	0.49	0.48	0.50	0.44
FreiHand (F)	0.15	0.40	0.72	0.37	0.35	0.43	0.35
HUMBI (H)	0.16	0.36	0.18	0.50	0.43	0.47	0.41
Average	0.30	0.44	0.36	**0.46**	**0.42**	**0.48**	**0.41**

Table 5: Cross-data evaluation results of 3D hand keypoint prediction. Metric is AUC of PCK calculated over an error range of 0-20 mm.

Monocular 3D Hand Mesh Prediction We compare HUMBI Hand with synthetic ObMan [21] dataset. We use a recent regression network [77] that outputs the hand mesh shape and camera pose with minor modifications, e.g., we change the size of the latent coefficient and the hand mesh decoder to the ones from the MANO hand model. We train and evaluate the network based on the reprojection error with weak perspective projection model. The results are summarized in Table 6. Due to the domain gap between the real and synthetic data, the prediction accuracy of the network trained with synthetic data is largely degraded on the real data. However, by combining two datasets, the performance is highly improved (even better than intra-data evaluation), e.g., ObMan+HUMBI can outperform ObMan and HUMBI 0.3 and 1.7 pixels, respectively.

4.4. Body

Benchmark Datasets We use four benchmark datasets: (1) Human3.6M [23] contains numerous 3D human poses of 11 actors/actresses measured by motion capture system with corresponding images from 4 cameras. (2) MPI-INF-3DHP [41] is 3D human pose estimation dataset, which contains both 3D and 2D pose labels as well as images covering both indoor and outdoor scenes. We use its test set containing 2,929 valid frames from 6 subjects. (3) UP-3D [33] is a 3D body mesh dataset providing ∼9K pairs of 3D body reconstruction and the associated 2D images. We use Human3.6M, MPI-INF-3DHP for body pose evaluation and UP-3D for body mesh evaluation.

Monocular 3D Body Pose Prediction To validate HUMBI body, we conduct a cross-data evaluation for the task of estimating 3D human pose from a single view image. We use a recent body pose detector [83]. We train and evaluate model trained by each dataset and model trained by a combination of HUMBI and each other dataset. By following the training protocol of [83], we use 2D landmark labels from MPII dataset [5] for a weak supervision. The results are summa-

Testing \ Training	ObMan	HUMBI	ObMan+HUMBI
ObMan	3.84±2.6	6.1±4.1	**3.5±2.4**
HUMBI	10.6±11.3	6.5±8.4	**4.8±5.8**

Table 6: The mean error of 3D hand mesh prediction for cross-data evaluation (unit: pixel).

Testing \ Training	H36M	MI3D	HUMBI	H36M +HUMBI	MI3D +HUMBI
H36M	0.562	0.362	0.434	0.551	0.437
MI3D	0.317	0.377	0.354	0.375	0.425
HUMBI	0.248	0.267	0.409	0.372	0.377
Average	0.376	0.335	**0.399**	**0.433**	**0.413**

Table 7: Cross-data evaluation results of 3D body keypoint prediction. Metric is AUC of PCK calculated over an error range of 0-150 mm.

Testing \ Training	UP-3D	HUMBI	UP-3D+HUMBI
UP-3D	22.7±18.6	49.4±0.09	**18.4±13.8**
HUMBI	26.0±19.7	14.5±6.6	**12.5±8.4**

Table 8: The mean error of 3D body mesh prediction for cross-data evaluation (unit: pixel).

rized in Table 7. We use area under PCK curve (AUC) in an error range of 0-150 mm as the metric. It show that HUMBI is more generalizable for predicting 3D body pose than Human3.6M and MPI-INF-3DHP (by a margin of 0.023 and 0.064 AUC). Moreover, HUMBI is complementary to each other dataset and the performance of model trained by another dataset alone is increased with HUMBI (by a margin of 0.057 and 0.078 AUC respectively).

Monocular 3D Body Mesh Prediction We compare the body mesh prediction accuracy using a recent CNN model trained on (1) HUMBI, (2) UP-3D, and (3) HUMBI+UP-3D. While we use [77] for the testing CNN model, recent monocular body reconstruction methods [1–4, 9, 20, 30, 35, 46, 48, 54, 58] can be alternative to test the generalization ability of HUMBI. The network decoder is modified to accommodate the differentiable SMPL parameter prediction. The reprojection error is used to supervise the network and to evaluate testing performance. The cross-data evaluation is summarized in Table 8. We observe that the network trained with HUMBI shows weak performance because of the lack of diversity of poses. However, it is highly complementary to other datasets as it provides various appearance from 107 viewpoints as shown in Figure 7.

4.5. Garment

We conduct camera-ablation study to evaluate how the number of cameras affect garment reconstruction quality. We incrementally reduce the number of cameras and measure the reconstruction accuracy and density. The reconstruction density is computed by the number of 3D points produced by multiview stereo [59]. The reconstruction accuracy metric is the closest point distance from the 3D garment surface reconstructed by shape-from-silhouette [34]. In both cases, the performance reaches to the optimal even without 107 cameras as shown in Figure 8, ensuring that our garment reconstruction is accurate (density: 90 cameras ≈ 107 cameras; accuracy: 60 cameras ≈ 107 cameras). The additional evaluations on the garment silhouette accuracy can be found in the Appendix.

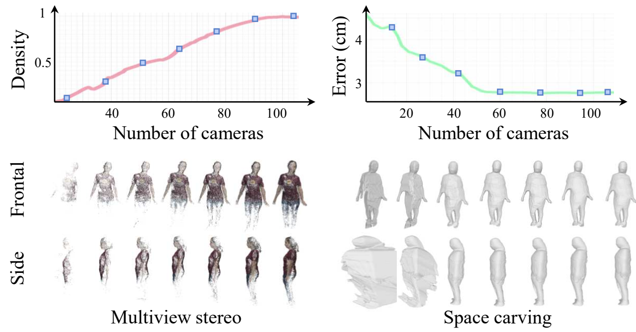


Figure 8: We conduct camera-ablation study to evaluate the accuracy of the garment reconstruction in terms of the density (left) and the accuracy (right).

5. Discussion

We present HUMBI dataset that is designed to facilitate high resolution pose- and view-specific appearance of human body expressions. Five elementary body expressions (gaze, face, hand, body, and garment) are captured by a dense camera array composed of 107 synchronized cameras. The dataset includes diverse activities of 772 distinctive subjects across gender, ethnicity, age, and physical condition. We use a 3D mesh model to represent the expressions where the view-dependent appearance is coordinated by its canonical atlas. Our evaluation shows that HUMBI outperforms existing datasets as modeling nearly exhaustive views and can be complementary to such datasets.

HUMBI is the first-of-its-kind dataset that attempts to span the general appearance of assorted people by pushing towards two extremes: views and subjects. This will provide a new opportunity to build a versatile model that generates photorealistic rendering for authentic telepresence. However, the impact of HUMBI will not be limited to appearance modeling, i.e., it can offer a novel multiview benchmark dataset for a stronger and generalizable reconstruction and recognition model specific to humans.

Acknowledgement

This work was partially supported by National Science Foundation (No.1846031 and 1919965), National Research Foundation of Korea, and Ministry of Science and ICT of Korea (No. 2020R1C1C1015260).

References

[1] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*. CVPR Spotlight Paper. 8

[3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video. In *3DV*, 2018. 8

[4] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 8

[5] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 7

[6] M. Armando, J.-S. Franco, and E. Boyer. Adaptive mesh texture for multi-view appearance modeling. In *3DV*, 2018. 1

[7] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. *SIGGRAPH*, 2010. 1

[8] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *TPAMI*, 2013. 2

[9] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *ICCV*, 2019. 8

[10] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *TPAMI*, 2003. 2, 3

[11] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *CVPR*, 2017. 2, 3

[12] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3D morphable models. *IJCV*. 3

[13] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, and T. Boubekeur. Markerless garment capture. In *TOG*, 2008. 2

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 3

[15] J. Y. B. R. M. A. Christian Zimmermann, Duygu Ceylan and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, 2019. 2, 3, 7

[16] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *TPAMI*, 2001. 1, 2

[17] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 2

[18] T. Fischer, H. Jin Chang, and Y. Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *ECCV*, 2018. 2, 3, 6

[19] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-PIE. *IVC*, 2009. 3

[20] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 2019. 8

[21] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 7

[22] 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. 3, 4, 6

[23] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2014. 2, 3, 7

[24] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 1

[25] H. Joo, H. S. Park, and Y. Sheikh. Map visibility estimation for large-scale dynamic 3d reconstruction. In *CVPR*, 2014. 1

[26] H. Joo, T. Simon, M. Cikara, and Y. Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *CVPR*, 2019. 1

[27] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 2017. 2, 3

[28] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 2

[29] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *CVPR*, 2016. 2

[30] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 8

[31] D. Knossow, R. Ronfard, and R. Horaud. Human motion tracking with a kinematic parameterization of extremal contours. *IJCV*, 2008. 2, 3

[32] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 2

[33] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 7

[34] A. Laurentini. The visual hull concept for silhouette-based image understanding. *TPAMI*, 1994. 4, 8

[35] V. Lazova, E. Insafutdinov, and G. Pons-Moll. 360-degree textures of people in clothing from a single image. In *3DV*, 2019. 8

[36] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012. 2

[37] G. Lin, A. Milan, C. Shen, and I. D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 4

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[39] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. *SIGGRAPH*, 2018. 1, 2, 4

[40] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *TOG*, 2015. 2, 4
[41] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017. 7
[42] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *AVBPA*, 1999. 2
[43] K. A. F. Mora, F. Monay, and J.-M. Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *ETRA*, 2014. 2, 3
[44] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, 2018. 2
[45] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *ICCV*, 2017. 2
[46] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018. 8
[47] S. Park and J. Hodgins. Capturing and animating skin deformation in human motion. *SIGGRAPH*, 2006. 3
[48] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 8
[49] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. *AVSS*, 2009. 3, 6
[50] A. Pentland. *Honest Signals: How They Shape Our World*. MIT Press, 2008. 1
[51] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *TOG*, 2017. 1, 3, 5
[52] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: A model of dynamic human shape in motion. *SIGGRAPH*, 2015. 1, 2, 3
[53] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH*, 2017. 2, 4
[54] Y. Rong, Z. Liu, C. Li, K. Cao, and C. C. Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *ICCV*, 2019. 8
[55] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *IVC*, 2016. 2
[56] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, 2013. 2
[57] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPRW*, 2013. 2
[58] S. Saito, , Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. 2019. 8
[59] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3, 4, 8
[60] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 2010. 2, 3
[61] T. Simon, H. Joo, I. A. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2, 3
[62] B. Smith, Q. Yin, S. Feiner, and S. Nayar. Gaze locking: Passive eye contact detection for humanobject interaction. In *UIST*, 2013. 2, 3
[63] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, 2007. 5
[64] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-bysynthesis for appearance-based 3d gaze estimation. In *CVPR*, 2014. 2, 3, 6
[65] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *CVPR*, 2015. 2
[66] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramana. Depth-based hand pose estimation: data, methods, and challenges. In *ICCV*, 2015. 2
[67] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *CVPR*, 2014. 2
[68] A. Tewari, M. Zollöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*, 2017. 2
[69] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *TOG*, 2014. 2, 3
[70] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2
[71] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 3
[72] T. von Marcard, G. Pons-Moll, and B. Rosenhahn. Human pose estimation from video and imus. *Transactions on Pattern Analysis and Machine Intelligence*, 2016. 3
[73] A. Wenger, A. Gardner, C. Tchou, J. Unger, T. Hawkins, and P. Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. *SIGGRAPH*, 2005. 1
[74] A. Wetzler, R. Slossberg, and R. Kimmel. Rule of thumb: Deep derotation for improved fingertip detection. In *BMVC*, 2016. 2, 3
[75] R. White, K. Crane, and D. A. Forsyth. Capturing and animating occluded cloth. In *TOG*, 2007. 2
[76] J. S. Yoon, Z. Li, and H. S. Park. 3d semantic trajectory reconstruction from 3d pixel continuum. *CVPR*, 2018. 5
[77] 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. 6, 7, 8
[78] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim. Big hand 2.2m benchmark: Hand pose data set and state of the art analysis. In *CVPR*, 2017. 2, 3

[79] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, 2017. 3
[80] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang. A hand pose tracking benchmark from stereo matching. In *ICIP*, 2017. 3, 7
[81] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, 2015. 2, 3, 6
[82] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *ICCVW*, 2013. 2
[83] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 7
[84] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 2, 6
[85] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 2
[86] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017. 2, 3, 7