# Supplementary Material
# AGORA: Avatars in Geography Optimized for Regression Analysis

Priyanka Patel[1] Chun-Hao P. Huang[1] Joachim Tesch[1] David T. Hoffmann[2,3]
Shashank Tripathi[1] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of Freiburg [3]Bosch Center for Artificial Intelligence
{ppatel, paul.huang, jtesch, dhoffmann, stripathi, black}@tuebingen.mpg.de

## 1. Method

Here we further explain the method section of the main paper in detail. We initialize the parameters of the SMPL-X [12] model with a multi-view fitting approach, followed by a refinement step that fits SMPL-X to the 3D scan as show in Fig. 1.

### 1.1. Multi-view Initialization.

The scans contain arbitrary poses, varied clothing, and people holding objects. This makes the automatic fitting of SMPL-X a challenge without a good initialization. We first center $S$ and render images of it from $C$ pre-defined virtual cameras. 2D landmarks are detected in each rendered image with [3] and we initialize the parameters with an approach that extends the single-view SMPLify-X fitting [12] to incorporate landmarks in multiview images.

SMPLify-X [12] takes one color image as input and optimizes the pose $\theta$, shape $\beta$ and facial expression $\psi$ of SMPL-X to match the observed 2D landmarks by minimizing the following objective:

$$E(\beta, \theta, \psi) = E_J + E_{\text{reg}}, \quad (1)$$

$$E_{\text{reg}} = \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_h} E_{\theta_h} + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\mathcal{E} E_\mathcal{E} + \lambda_\mathcal{C} E_\mathcal{C}, \quad (2)$$

where $E_J$ is the data term that penalizes differences between projected and observed landmarks, and $E_{\text{reg}}$ includes several regularization terms: $E_\alpha(\theta_b)$ penalizes strong bending of elbows and knees, while $E_C$ prevents mesh-intersections. $E_{\theta_b}(\theta_b)$, $E_{\theta_h}(\theta_h)$, $E_\beta(\beta)$, and $E_\mathcal{E}(\psi)$ are $L_2$ priors on the body pose, hand pose, body shape and facial expressions. $\lambda$'s denote weights for each respective term. We adapt Eq. 2 to take multi-view data with known camera parameters for each camera $c$: $E_{\text{mv}} = E_{\text{reg}} + \sum_{c=1}^{C} E_J^c$. Unlike in [12] where one needs to estimate camera translation first, here the intrinsics and extrinsics are given.

### 1.2. 2D+3D Refinement

To get the skin and cloth vertices from scan, we use segmentation masks provided by Renderpeople to group scan points into skin, clothing (including shoes), and the rest (hair and objects). Since we do not have segmentation masks from other vendors, we generate them using Graphonomy [6]. Graphonomy provides human parts segmentation given an image with labels for cloth as well as body parts.

**Skin-Cloth Segmentation.** For each rendered multi-view image of a scan, Graphonomy outputs segmentation masks for different body parts and types of clothing. We group these into 3 labels: skin, cloth and other. Given the known cameras, we project visible vertices into the images and give them the corresponding label. Aggregating labels across all views gives a us a probability of each vertex being skin, clothing, or other. For Renderpeople scans, the probability is either 1 or 0 as we have segmentation masks. Similar to [16], we define energy terms $E_{\text{skin}}$ and $E_{\text{cloth}}$ for skin and clothing scan vertices, respectively.

Here we explain in detail, the two optimization terms, $E_{\text{skin}}$ and $E_{\text{cloth}}$ used in 2D+3D refinement. Please refer to the main paper for the full equation.

**Skin term**. For each scan vertex $s \in S$ we find the point on the closest model triangle. We minimize the point-to-surface distance between them weighted by probability $p \in P_{skin}$. Here $P_{skin}$ is the probability of the vertex $s$ belonging to skin calculated using Graphonomy [6].

$$E_{\text{skin}}(\beta, \theta, \psi) = \sum_{s \in \text{S}, p \in P_{\text{skin}}} \rho\left(p \cdot \text{dist}(s, M(\beta, \theta, \psi))\right), \quad (3)$$

where $\text{dist}(\cdot)$ represents the distance of the closest point on the model $M(\beta, \theta, \psi)$ surface from the scan vertex $s$. $\rho(\cdot)$ is Geman-McClure robust error function [5] that prevents outliers from contributing too much in the energy. Of course, we start with the initialization from Sec. 1.1.

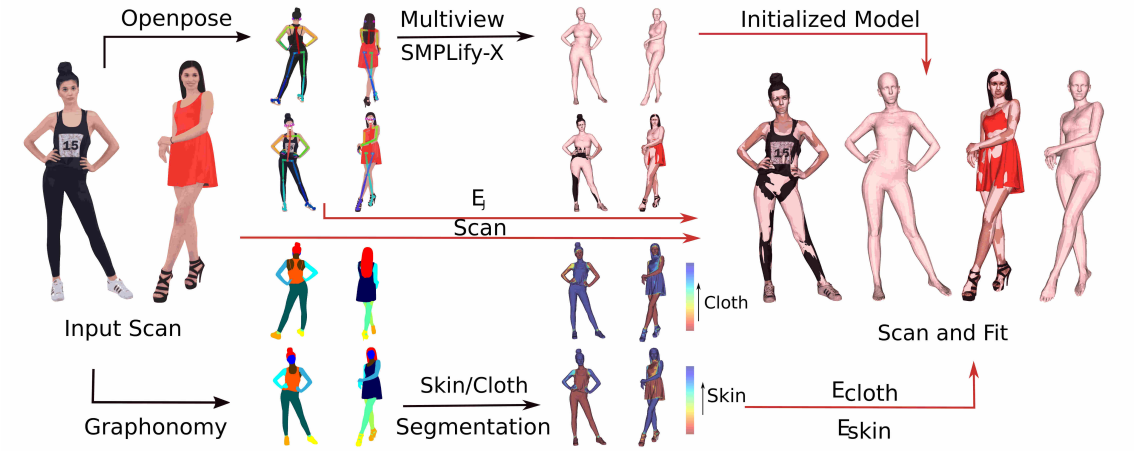


Figure 1. **SMPL-X fitting to scans.** Keypoints and body part segmentation across multiple rendered views are generated using OpenPose and Graphonomy. Multiview SMPLify-X initializes the model in proper pose. Shape is further refined using 2D+3D refinement (see text).

**Clothing term**. The goal of $E_{\text{cloth}}$ is to prevent clothing scan points from penetrating inside the model while keeping the model close to the scan, so that the body does not shrink.

Each scan point is further classified into two categories: points penetrating the body model $S_P$ and points outside the body model $S_O$. We get the probability of each scan vertex being cloth $P_{cloth}$ from Graphonomy [6]. $P_{cp}$ and $P_{co}$ are the corresponding cloth probability values for $S_P$ and $S_O$ taken from $P_{Cloth}$. For $s_i \in S_P$ we penalize the distance with weight $\lambda$, while for $s_i \in S_O$ we use again Geman-McClure function to accommodate loose clothes like skirts, saris, bath robes, etc. We weight the dist$(\cdot)$ with the corresponding probability values $P_{cp}$ and $P_{co}$. Specifically:

$$E_{\text{cloth}}(\beta, \theta; \psi) = \sum_{p \in P_{co}, s \in S_O} \rho\left(p \cdot \text{dist}(s, M(\beta, \theta, \psi))\right) + \lambda\Big( \sum_{p \in P_{cp}, s \in S_P} p \cdot \text{dist}(s, M(\beta, \theta, \psi))\Big), \quad (4)$$

where we do not optimize facial expression, $\psi$, because it is not covered by clothing and where dist$(\cdot)$ and $\rho$ are the same as in Eq. 3.

Since the initialization in Sec. 1.1 is already close, the classification of $S_P$ and $S_O$ can be approximated as follows. Each vertex $s_i \in S$ has a point $m_i$ on the nearest triangle of the model with a corresponding normal $n_i$. We define a displacement vector $d_i = m_i - s_i$ and identify $S_P$ if the inner product of $d_i$ and $n_i$ is greater than 0, otherwise we consider it $S_O$.

### 1.3. Child scans fitting

As described in Sec. 3.2 of the main paper, we fit 257 children scans by using a template that is an interpolation of adult SMPL-X template and SMIL infant template [7].

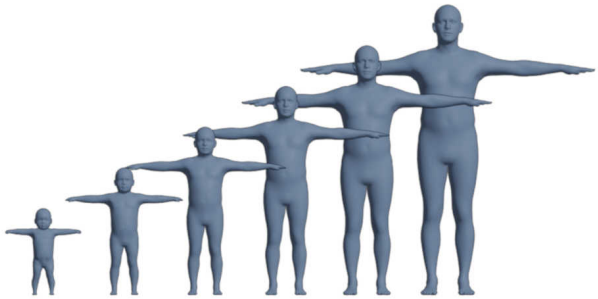

Figure 2. SMIL-X and adult male SMPL-X template interpolation.

Fig. 2 shows how varying interpolation coefficient gives us approximate template from different age group.

## 2. AGORA

**AGORA Statistics.** We provide the dataset distribution across various attributes i.e. age, ethnicity and gender for AGORA ground truth scans in Fig. 3. AGORA has evenly distributed gender and a varied range of age and ethnicity. To create this distribution, we use gender, age and ethnicity information provided by Renderpeople [2]. For other vendors since no age and ethnicity information was given we label them with the help of Amazon Mechanical Turks [1]. We recruit 5 different subjects with $>$ 5000 HITs approved and an approval rate $>$97%. We ask them to classify the rendered image of the scan into predefined categories of ethnicity and age (including Unknown). If there is a majority vote, we label the scan with the respective category. If there is a tie, we resolve it by ourselves by selecting the best estimate or by selecting Unknown. For others, we marked them Unknown. We label all the scans with gender ourselves.

**AGORA Dataset.** Fig. 6 provides more examples of our

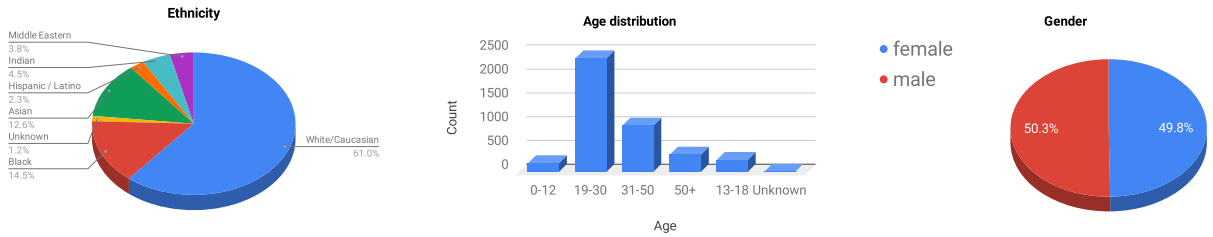


Figure 3. Breakdown of AGORA dataset in ethnicity, age, and gender.

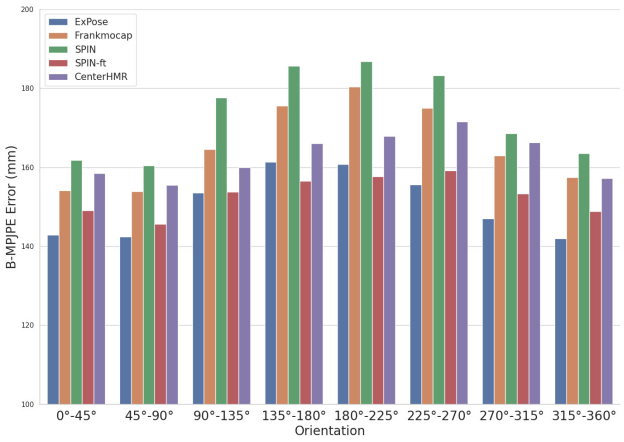


Figure 4. Error in varied orientations relative to the camera. $0°$ corresponds to facing the camera. Evaluated on BFH subset of AGORA for 22 SMPL-X and 24 SMPL joints.

dataset. From left to right we show the RGB image, the RGB image with ground truth SMPL-X fits and segmentation masks. The 3D scenes (row 1-4) lead to challenging environmental occlusion, as can be seen in the segmentation masks on the right. Row 5 shows a example from the easy-split experiment described in Sec. 3.2. We also render individual subject masks as though there is no occlusion. See Fig. 5 for examples. We use these masks to determine how much a person is occluded.

## 3. Additional Analysis on Baseline Experiment

Here we provide details for the Evaluation Protocol as well as additional analysis of the baselines.

### 3.1. Evaluation Protocol.

In the following we describe our evaluation protocol in detail. A less detailed description with visualization can be found in the main paper in Section 4.3. We evaluate the following methods: ExPose [4], FrankMocap [13], SMPLify-X [12], HMR [10], SPIN [11], EFT [9], and CenterHMR [14]. which collectively provide a good picture of the current SOTA. Our protocol can be split into four parts: 1) Detection, 2) 3D pose and shape estimation, 3) matching of predictions to ground truth and 4) computing errors. In the following we will explain each component.

**Detection.** All the methods we evaluate require the person to first be localized. To obtain person detections we run OpenPose [3] on the input image. OpenPose requires a large number of settings. To select these settings we draw inspiration from the maximum accuracy setting as reported on the OpenPose GitHub page[1]. However, these settings have huge memory requirements on the GPU. Thus, we modify the settings such that OpenPose can run on a common GPU with 12 GB of memory. We end up with the following settings: We scale the larger side of the input images to 272 pixels while keeping the aspect ratio fixed. We use two scale processing for the body keypoints with a scale gap of 0.25. We run the face detection network with default settings. Note that only SMPLify-X makes use of hand and face detections, however, we use the same settings for all the evaluated methods to reduce influence of the keypoint detection on the results. For CenterHMR [14] we directly use the entire image as input without any cropping.

**3D pose and shape estimation.** For each OpenPose detection we run all the methods. All methods in our experiment except [14] require either 2D keypoints or tight crops around the detected person as input. For ExPose [4], FrankMocap [13],HMR [10] and SPIN [11] we use their demo code to generate crops from OpenPose detections. SMPLify-X [12] operates directly on OpenPose keypoints and no pre-processing is needed. EFT considers both the image feature of the crop and keypoints. For CenterHMR [14] we directly use the entire image as input for 3D pose and shape estimation.

**Matching predictions to ground truth.** Let $M$ be the set of predicted meshes and $N$ be the set of ground truth humans. To match the predictions of the methods to ground truth, we project the 3D keypoints of the estimated SMPL-body to the image plane. To this end, the camera parameters as assumed or estimated by the method are used. Similarly, we project the ground truth 3D keypoint to the image plane using the ground truth camera parameters. We compute the 2D joint error for all combinations of $m \in M$ and $n \in N$

[1] https://github.com/CMU-Perceptual-Computing-Lab/openpose

Figure 5. Person segmentation masks. Left: color images. Middle: full masks. Right: individual masks rendered with no occlusions.

and match them based on minimal 2D keypoint error.

**False positives and false negatives.** The predictions are often noisy leading to *false positives* and *false negatives*, as shown in Fig. 3 of the main paper. It is crucial to detect false positives to avoid that a false positive is incorrectly matched to ground truth, resulting in large errors, which might distort the results. To detect false positives we construct 2D axis-aligned bounding boxes (AABB) for each prediction and ground truth based on the 2D keypoints. Before matching a prediction with ground truth we compute the intersection over union (IoU) for this pair. If IoU $< \tau$ and no other possible match exists for a given prediction it is considered a false positive. We choose our threshold $\tau = 0.1$, such that predictions which have a large distance in 2D are not considered to match, but small enough to ensure that differences in the scale of prediction and ground truth do not lead to erroneous classification as a false positive. Finally, each unmatched ground truth body is considered as a false negative.

**Computing errors.** We compute the errors using different metrics, as described in Section 4.2 in the main paper.

### 3.2. Evaluation of Methods

A qualitative comparison of different methods is shown in Fig. 7. In Sec. 5 of the main paper we analyse the dependence of errors with respect to occlusion and distance from the center of the image. Here, we analyze the dependence of errors on the body orientation.

**Orientation.** We concentrate particularly on the yaw rotation with respect to the camera, and report the error in Fig. 4, where $0^\circ$ corresponds to facing the camera. We observe that the error grows as the yaw angle increases, reaching the peak around $180^\circ$ and then decreases. This suggests that the current 3D human-pose-and-shape methods perform worst when subjects are not facing the camera.

**From scratch training.**

| Models | 3DPW (14 joints) | | 3DPW (24 joints) | |
|---|---|---|---|---|
| | MPJPE ↓ | PA-MPJPE ↓ | MPJPE ↓ | PA-MPJPE ↓ |
| Human3.6M [8] | 311.3 | 162.1 | 286.2 | 178.1 |
| $[\text{MPII+LSPet+COCO}]_{\text{EFT}}$ [9] | 125.0 | 77.4 | 121.9 | 86.1 |
| AGORA | 147.4 | 81.0 | 141.3 | 88.8 |

Table 1. Training SPIN from scratch with Human3.6M vs. EFT vs. AGORA.

In Table 1 we report results for SPIN trained from scratch using different datasets. For AGORA training, we report on-par PA-MPJPE compared to $[\text{MPII+LSPet+COCO}]_{\text{EFT}}$ but much better results than for training from scratch with Human3.6M [8]. These experiments suggest that the AGORA training set is sufficiently realistic and large to support both finetuning and from-scratch training. Pre-trained weights on AGORA will be made available for research purposes.

**Easy split experiment.** To further show that the images in AGORA are comparable in complexity to natural images, we create an easy test set of approximately 400 images as sanity check. Each image consists of only two people in the easy split, potentially with some minor occlusion (5th row of Fig. 6). Despite having two person per image and minor occlusion, easy test set is still more challenging than 3DPW [15] because of varied lighting and complex clothing. We evaluate FrankMocap [13] and ExPose on this set and report 118.0 and 111.6mm B-MPJPE error for 22 SMPL-X joints. ExPose error is comparable to the reported error on 3DPW in original work [4] as 93.4. Since, FrankMocap only shows qualitative results, we can't compare the quantitative evaluation with original work. The easy-split experiment along with the finetuning experiment described in main paper suggest that our synthetic images are on par with natural images.

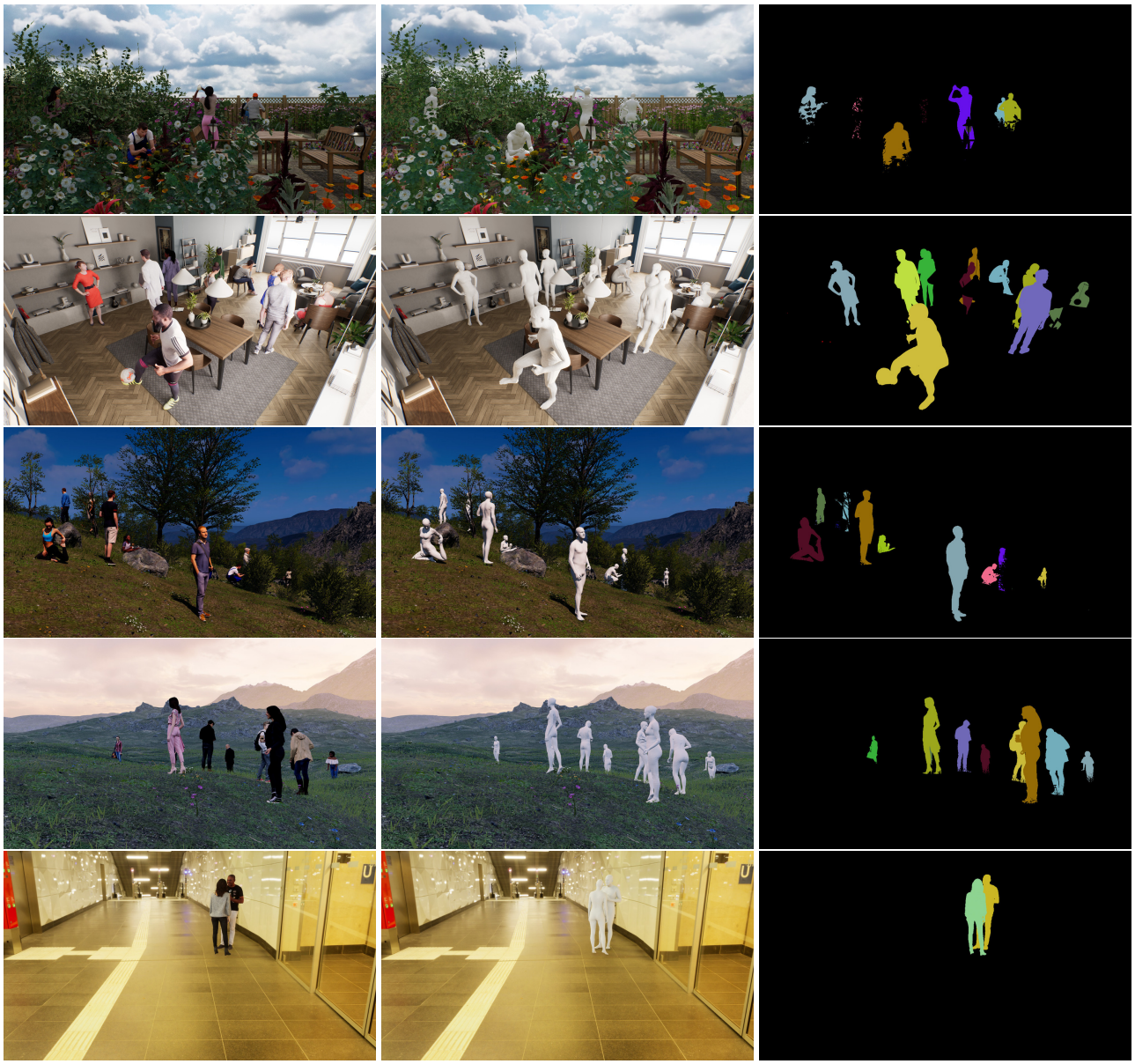

Figure 6. Examples images from the AGORA dataset.

## References


[1] *mturk*, 2020. https://www.mturk.com/. 2

[2] *Renderpeople*, 2020. https://renderpeople.com. 2

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021. 1, 3

[4] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40, 2020. 3, 4

[5] Stuart Geman. Statistical methods for tomographic image reconstruction. *Bull. Int. Stat. Inst*, 4:5–21, 1987. 1

[6] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7450–7459, 2019. 1, 2

[7] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. Learning an infant body model from RGB-D data for accurate full body motion analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 792–800, 2018. 2

[8] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 4

[9] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020. 3, 4

[10] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 3

[11] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision*, pages 2252–2261, 2019. 3

[12] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 1, 3

[13] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3D hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 3, 4

[14] Yu Sun, Qian Bao, Wu Liu, Yili Fu, and Tao Mei. CenterHMR: a bottom-up single-shot method for multi-person 3D mesh recovery from a single image. *arXiv preprint arXiv:2008.12272*, 2020. 3

[15] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision*, pages 614–631, 2018. 4

[16] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 1

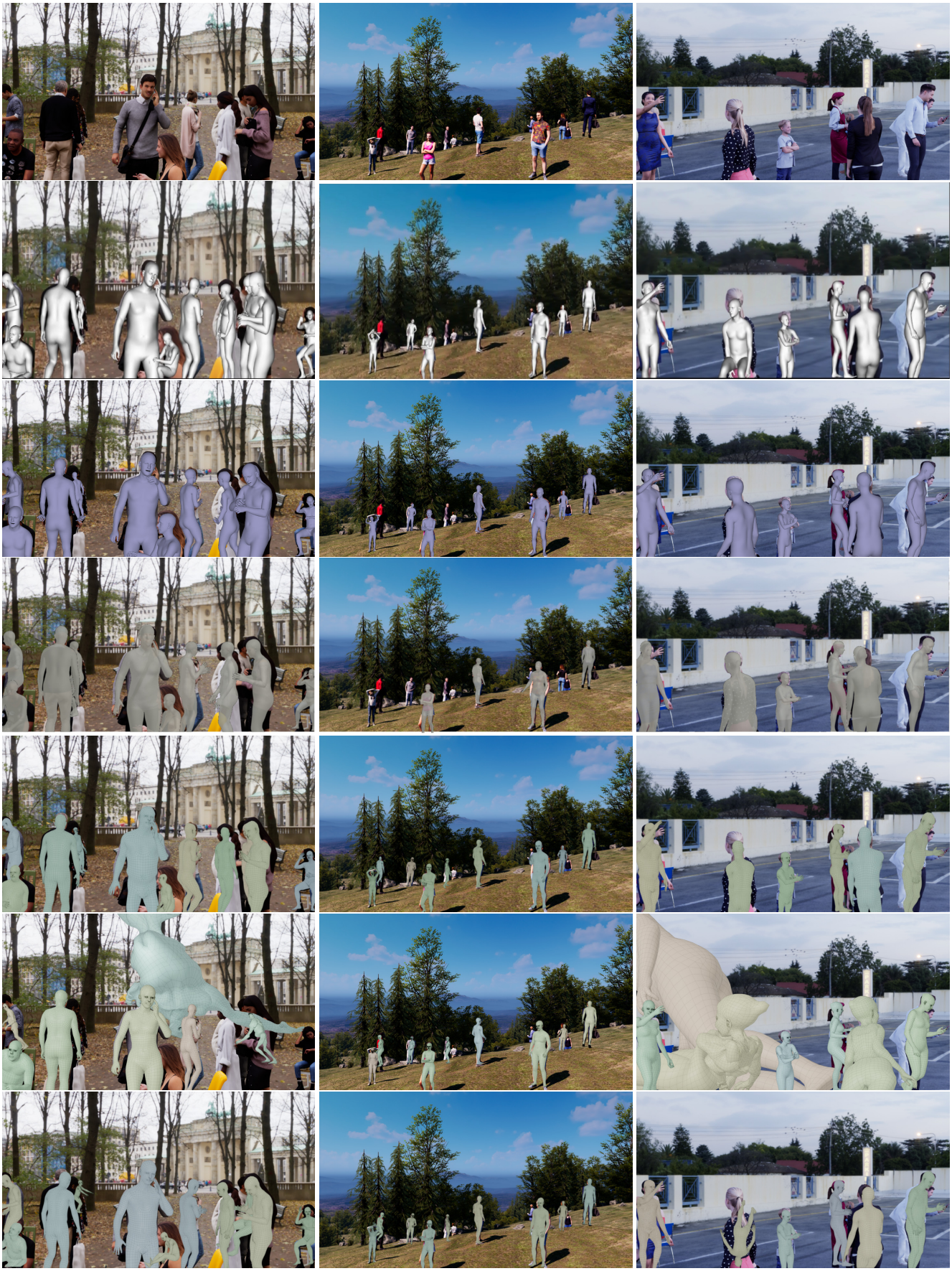

Figure 7. Method evaluation. RGB images (row 1), FrankMocap (row 2), ExPose (row 3), CenterHMR (row 4), HMR (row 5), SMPLify-X (row 6) and SPIN (row 7).