Monocular Expressive Body Regression through Body-Driven Attention

Vasileios Choutas[1,2], Georgios Pavlakos[3], Timo Bolkart[1], Dimitrios Tzionas[1], and Michael J. Black[1]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] Max Planck ETH Center for Learning Systems
[3] University of Pennsylvania, Philadelphia, USA
{vchoutas, gpavlakos, tbolkart, dtzionas, black}@tuebingen.mpg.de

Abstract. To understand how people look, interact, or perform tasks, we need to quickly and accurately capture their 3D body, face, and hands *together* from an RGB image. Most existing methods focus only on parts of the body. A few recent approaches reconstruct full expressive 3D humans from images using 3D body models that include the face and hands. These methods are optimization-based and thus slow, prone to local optima, and require 2D keypoints as input. We address these limitations by introducing *ExPose* (EXpressive POse and Shape rEgression), which directly regresses the body, face, and hands, in SMPL-X format, from an RGB image. This is a hard problem due to the high dimensionality of the body and the lack of expressive training data. Additionally, hands and faces are much smaller than the body, occupying very few image pixels. This makes hand and face estimation hard when body images are downscaled for neural networks. We make three main contributions. First, we account for the lack of training data by curating a *dataset* of SMPL-X fits on in-the-wild images. Second, we observe that body estimation localizes the face and hands reasonably well. We introduce *body-driven attention* for face and hand regions in the original image to extract higher-resolution crops that are fed to dedicated refinement modules. Third, these modules exploit *part-specific knowledge* from existing face- and hand-only datasets. ExPose estimates expressive 3D humans more accurately than existing optimization methods at a small fraction of the computational cost. Our data, model and code are available for research at https://expose.is.tue.mpg.de.

1 Introduction

A long term goal of computer vision is to understand humans and their behavior in everyday scenarios using only images. Are they happy or sad? How do they interact with each other and the physical world? What are their intentions? To answer such difficult questions, we first need to *quickly* and *accurately* reconstruct their 3D body, face and hands *together* from a single RGB image. This is very challenging. As a result, the community has broken the problem into pieces with much of the work focused on estimating either the main body [18, 62, 78], the face [106] or the hands [14, 87, 97] separately.

Fig. 1: *Left:* Full-body RGB images of people contain many more pixels on the body than on the face or hands. *Middle:* Images are typically downsized (e.g. to 256×256 px) for use in neural networks. This resolution is fine for the body but the hands and face suffer from low resolution. Our model (Figure 2) uses *body-driven attention* to restore the lost information for hands and faces from the original image, feeding it to dedicated refinement modules. *Right:* These modules give more expressive hands and faces, by exploiting *part-specific knowledge* learned from higher quality hand-only [105] and face-only [41] datasets; green meshes show example part-specific training data.

Only recent advances have made the problem tractable in its full complexity. Early methods estimate 2D joints and features [10, 29] for the body, face and hands. However, this is not enough. It is the skin surface that describes important aspects of humans, e.g. what their precise 3D shape is, whether they are smiling, gesturing or holding something. For this reason, strong statistical parametric models for expressive 3D humans were introduced, namely Adam [38], SMPL-X [67] and recently GHUM/GHUML [96]. Such models are attractive because they facilitate reconstruction even from ambiguous data, working as a strong prior.

There exist three methods that estimate full expressive 3D humans from an RGB image [67, 95, 96], using SMPL-X, Adam and GHUM/GHUML respectively. These methods are based on optimization, therefore they are slow, prone to local optima, and rely on heuristics for initialization. These issues significantly limit the applicability of these methods. In contrast, recent body-only methods [39, 46] directly regress 3D SMPL bodies quickly and relatively reliably directly from an RGB image.

Here we present a *fast* and *accurate* model that reconstructs full *expressive* 3D humans, by estimating SMPL-X parameters directly from an RGB image. This is a hard problem and we show that it is not easily solved by extending SMPL neural-network regressors to SMPL-X for several reasons. First, SMPL-X is a much higher dimensional model than SMPL. Second, there exists no large in-the-wild dataset with SMPL-X annotations for training. Third, the face and hands are often blurry and occluded in images. At any given image resolution, they also occupy many fewer pixels than the body, making them low resolution. Fourth, for technical reasons, full body images are typically downscaled for input

to neural networks [48], e.g. to 256×256 pixels. As shown in Figure 1, this results in even lower resolution for the hands and face, making inference difficult.

Our model and training method, shown in Figure 2, tackles all these challenges. We account for data scarcity by introducing a new dataset with paired in-the-wild images and SMPL-X annotations. To this end, we employ several standard in-the-wild body datasets [3, 35, 36, 57] and fit SMPL-X to them with SMPLify-X [67]. We semi-automatically curate these fits to keep only the good ones as pseudo ground-truth. We then train a model that regresses SMPL-X parameters from an RGB image, similar to [39]. However, this only estimates rough hand and face configurations, due to the problems described above. We observe that the main body is estimated well, on par with [39, 46], providing good rough localization for the face and hands. We use this for *body-driven attention* and focus the network back on the *original* non-downscaled image for the face and hands. We retrieve high-resolution information for these regions and feed this to dedicated *refinement* modules. These modules act as an *expressivity boost* by distilling *part-specific knowledge* from high-quality hand-only [105] and face-only [58] datasets. Finally, the independent components are fine-tuned jointly end-to-end, so that the part networks can benefit from the full-body initialization.

We call the final model ExPose (EXpressive POse and Shape rEgression). ExPose is as accurate as existing optimization-based methods [67] for estimating expressive 3D humans, while running two orders of magnitude faster. Our data, model and code are available for research at https://expose.is.tue.mpg.de.

2 Related Work

Human Modeling: Modeling and capturing the whole human body is a challenging problem. To make it tractable, the community has studied the body, face and hands separately, in a divide-and-conquer fashion. For the human ***face***, the seminal work of Blanz and Vetter [6] introduces the first 3D morphable model. Since then, numerous works (see [13]) propose more powerful face models and methods to infer their parameters. For human ***hands*** the number of models is limited, with Khamis et al. [42] learning a model of hand shape variation from depth images, while Romero et al. [72] learn a parametric hand model with both a rich shape and pose space from 3D hand scans. For the human ***body***, the introduction of the CAESAR dataset [70] enables the creation of models that disentangle shape and pose, such as SCAPE [4] and SMPL [59], to name a few. However, these models have a neutral face and the hands are non-articulated. In contrast, Adam [38] and SMPL-X [67] are the first models that represent the body, face and hands jointly. Adam lacks the pose-dependent blendshapes of SMPL and the released version does not include a face model. The GHUM [96] model is similar to SMPL-X but is not publicly available at the time of writing.

Human Pose Estimation: Often pose estimation is posed as the estimation of 2D or 3D keypoints, corresponding to anatomical joints or landmarks [9, 10, 82]. In contrast, recent advances use richer representations of the 3D body surface in the form of parametric [7, 39, 65, 69] or non-parametric [47, 75, 92] models.

To estimate ***bodies*** from images, many methods break the problem down into stages. First, they estimate some intermediate representation such as 2D joints [7, 20, 21, 30, 39, 61, 69, 81, 91, 101], silhouettes [1, 30, 69], part labels [65, 74] or dense correspondences [23, 73]. Then, they reconstruct the body pose out of this proxy information, by either using it in the data term of an optimized energy function [7, 30, 98] or "lifting" it using a trained regressor [39, 61, 65, 69, 91]. Due to ambiguities in lifting 2D to 3D, such methods use various priors for regularization, such as known limb lengths [51], a pose prior for joint angle limits [2], or a statistical body model [7, 30, 65, 69] like SMPL [59]. The above 2D proxy representations have the advantage that annotation for them is readily available. Their disadvantage is that the eventual regressor does not get to exploit the original image pixels and errors made by the proxy task cannot be overcome.

Other methods predict 3D pose directly from RGB pixels. Intuitively, they have to learn a harder mapping, but they avoid information bottlenecks and additional sources of error. Most methods infer 3D body joints [53, 68, 85, 86, 90], parametric methods estimate model parameters [39, 40, 46], while non-parametric methods estimate 3D meshes [47], depth maps [17, 83] voxels [92, 102] or distance fields [75, 76]. Datasets of paired indoor images and MoCap data [31, 80] allow supervised training, but may not generalize to in-the-wild data. To account for this, Rogez and Schmid [71] augment these datasets by overlaying synthetic 3D humans, while Kanazawa et al. [39] include in-the-wild datasets [3, 35, 36, 57] and employ a re-projection loss on their 2D joint annotations for weak supervision.

Similar observations can be made in the human hand and face literature. For ***hands***, there has been a lot of work on RGB-D data [97], and more recent interest in monocular RGB [5, 8, 24, 26, 32, 50, 63, 89, 104]. Some of the non-parametric methods estimate 3D joints [32, 63, 89, 104], while others estimate 3D meshes [19, 49]. Parametric models [5, 8, 26, 50, 100] estimate configurations of statistical models like MANO [72] or a graph morphable model [50]. For ***faces***, 3D reconstruction and tracking has a long history. We refer the reader to a recent comprehensive survey [106].

Attention for Human Pose Estimation: In the context of human pose estimation, attention is often used to improve prediction accuracy. Successful architectures for 2D pose estimation, like Convolutional Pose Machines [93] and Stacked Hourglass [64] include a series of processing stages, where the intermediate pose predictions in the form of heatmaps are used as input to the following stages. This informs the network of early predictions and guides its attention to relevant image pixels. Chu et al. [12] build explicit attention maps, at a global and part-specific level, driving the model to focus on regions of interest. Instead of predicting attention maps, our approach uses the initial body mesh prediction to define the areas of attention for hands- and face-specific processing networks. A similar practice is used by OpenPose [10], where arm keypoints are used to estimate hand bounding boxes, in a heuristic manner. Additionally, for HoloPose [22], body keypoints are used to pool part-specific features from the image.

A critical difference of ExPose is that, instead of simply pooling already computed features, we also process the region of interest at higher resolution, to

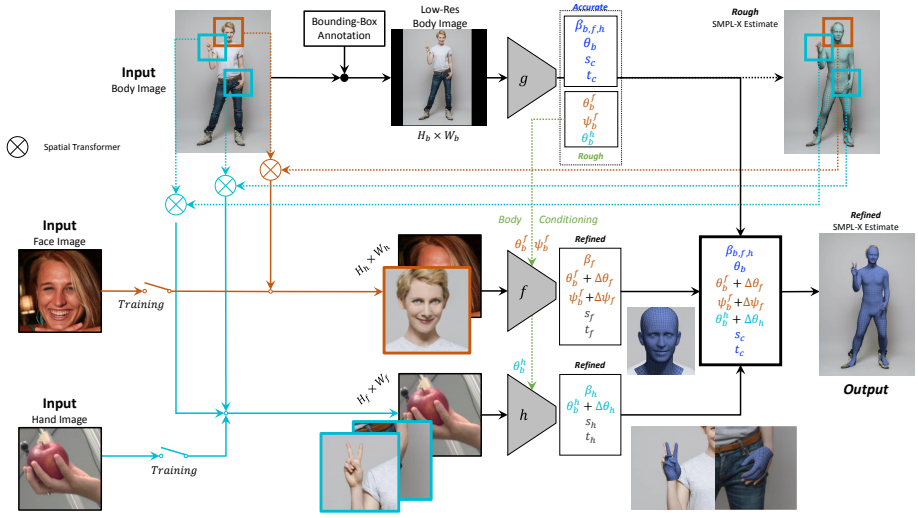

Fig. 2: An image of the body is extracted using a bounding box from the full resolution image and fed to a neural network $g(\cdot)$, that predicts body pose $\boldsymbol{\theta}_b$, hand pose $\boldsymbol{\theta}_h$, facial pose $\boldsymbol{\theta}_\text{f}$, shape $\boldsymbol{\beta}$, expression $\boldsymbol{\psi}$, camera scale s and translation $\boldsymbol{t}$. Face and hand images are extracted from the original resolution image using bilinear interpolation. These are fed to part specific sub-networks $f(\cdot)$ and $h(\cdot)$ respectively to produce the final estimates for the face and hand parameters. During training the part specific networks can also receive hand and face only data for extra supervision.

capture more subtle face and hand details. In related work, Chandran et al. [11] use a low resolution proxy image to detect facial landmarks and extract high resolution crops that are used to refine facial landmark predictions.

Expressive Human Estimation: Since expressive parametric models of the human body have only recently been introduced [38, 67, 72, 96], there are only a few methods to reconstruct their parameters. Joo et al. [38] present an early approach, but rely on an extended multi-view setup. More recently, Xiang et al. [95], Pavlakos et al. [67] and Xu et al. [96] use a single image to recover Adam, SMPL-X and GHUM parameters respectively, using optimization-based approaches. This type of inference can be slow and may fail in the presence of noisy feature detections. In contrast, we present the first regression approach for expressive monocular capture and show that it is both more accurate and significantly faster than prior work.

3 Method

3.1 3D Body Representation

To represent the human body, we use SMPL-X [67], a generative model that captures shape variation, limb articulation and facial expressions across a human population. It is learned from a collection of registered 3D body, hand and face scans of people with different sizes, nationalities and genders. The shape, $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and expression, $\boldsymbol{\psi} \in \mathbb{R}^{10}$, are described by 10 coefficients from the corresponding PCA spaces. The articulation of the limbs, the hands and the face is modeled by the pose vector $\boldsymbol{\theta} \in \mathbb{R}^{J \times D}$, where D is the rotation representation dimension, e.g. $D = 3$ if we select axis-angles, which describes the relative rotations of the $J = 53$ major joints. These joints include 22 main body joints, 1 for the jaw, and 15 joints per hand for the fingers. SMPL-X is a differentiable function $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$, that produces a 3D mesh $M = (V, F)$ for the human body, with $N = 10475$ vertices $V \in \mathbb{R}^{(N \times 3)}$ and triangular faces F. The surface of the articulated body is obtained by linear blend skinning driven by a rigged skeleton, defined by the above joints. Following the notation of [39] we denote posed joints with $X(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{J \times 3}$. The final set of SMPL-X parameters is the vector $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}\} \in \mathbb{R}^{338}$, as we choose to represent the pose parameters $\boldsymbol{\theta}$ using the representaion of Zhou et al. [103] with $D = 6$.

3.2 Body-driven Attention

Instead of attempting to regress body, hand and face parameters from a low resolution image crop we design an attentive architecture that uses the structure of the body and the full resolution of the image I. Given a bounding box of the body, we extract an image I_b, using an affine transformation $T_b \in \mathbb{R}^{2 \times 3}$, from the high-res image I. The body crop I_b is fed to a neural network g, similar to [39], to produce a first set of SMPL-X parameters Θ_b and weak-perspective camera scale $s_b \in \mathbb{R}$ and translation $\boldsymbol{t}_b \in \mathbb{R}^2$. After posing the model and recovering the posed joints X, we project them on the image:

$$\boldsymbol{x} = s(\Pi(X) + \boldsymbol{t}) \tag{1}$$

where Π is an orthographic projection. We then compute a bounding box for each hand and the face, from the corresponding subsets of projected 2D joints, $\boldsymbol{x}_h$ and $\boldsymbol{x}_f$. Let $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$ be the top left and bottom right points for a part, computed from the respective joints. The bounding box center is equal to $\boldsymbol{c} = \left(\frac{x_{\min}+x_{\max}}{2}, \frac{y_{\min}+y_{\max}}{2}\right)$, and its size is $b_s = 2 \cdot \max(x_{\max} - x_{\min}, y_{\max} - y_{\min})$. Using these boxes, we compute affine transformations T_h, $T_f \in \mathbb{R}^{2 \times 3}$ to extract higher resolution hand and faces images using spatial transformers (ST) [33]:

$$I_h = \mathrm{ST}\left(I; T_h\right), \; I_f = \mathrm{ST}\left(I; T_f\right). \tag{2}$$

The hand I_h and face I_f images are fed to a hand network h and a face network f, to refine the respective parameter predictions. Hand parameters $\boldsymbol{\theta}_h$ include

the orientation of the wrist $\boldsymbol{\theta}^{\text{wrist}}$ and finger articulation $\boldsymbol{\theta}^{\text{fingers}}$, while face parameters contain the expression coefficients $\boldsymbol{\psi}_f$ and facial pose $\boldsymbol{\theta}_f$, which is just the rotation of the jaw. We refine the parameters from by body by predicting offsets for each of the parameters and condition the part specific networks on the corresponding body parameters:

$$[\Delta\boldsymbol{\theta}^{\text{wrist}}, \Delta\boldsymbol{\theta}^{\text{fingers}}] = h\left(I_h; \boldsymbol{\theta}_b^{\text{wrist}}, \boldsymbol{\theta}_b^{\text{fingers}}\right), \; [\Delta\boldsymbol{\theta}_{\text{f}}, \Delta\boldsymbol{\psi}] = f\left(I_f; \boldsymbol{\theta}_b^{\text{f}}, \boldsymbol{\psi}_b\right) \tag{3}$$

where $\boldsymbol{\theta}_b^{\text{wrist}}$, $\boldsymbol{\theta}_b^{\text{fingers}}$, $\boldsymbol{\theta}_b^{\text{f}}$, $\boldsymbol{\psi}_b$ are the wrist pose, finger pose, facial pose and expression predicted by $g(\cdot)$. The hand and head sub-networks also produce a set of weak-perspective camera parameters $\{s_{\text{h}}, \boldsymbol{t}_{\text{h}}\}$, $\{s_{\text{f}}, \boldsymbol{t}_{\text{f}}\}$ that align the predicted 3D meshes to their respective images I_h and I_f. The final hand and face predictions are then equal to:

$$\boldsymbol{\theta}_h = \left[\boldsymbol{\theta}^{\text{wrist}}, \boldsymbol{\theta}^{\text{fingers}}\right] = \left[\boldsymbol{\theta}_b^{\text{wrist}}, \boldsymbol{\theta}_b^{\text{fingers}}\right] + [\Delta\boldsymbol{\theta}_{\text{wrist}}, \Delta\boldsymbol{\theta}_{\text{fingers}}] \tag{4}$$

$$[\boldsymbol{\psi}, \boldsymbol{\theta}_f] = \left[\boldsymbol{\psi}_b, \boldsymbol{\theta}_b^f\right] + [\Delta\boldsymbol{\psi}, \Delta\boldsymbol{\theta}_f]\,. \tag{5}$$

With this approach we can utilize the full resolution of the original image I to overcome the small pixel resolution of the hands and face in the body image I_b. Another significant advantage is that we are able to leverage hand- and face-only data to supplement the training of the hand and face sub-networks. A detailed visualization of the prediction process can be seen in Figure 2. The loss function used to train the model is a combination of terms for the body, the hands and the face. We train the body network using a combination of a 2D re-projection loss, 3D joint errors and a loss on the parameters Θ, when available. All variables with a hat denote ground-truth quantities.

$$L = L_{\text{body}} + L_{\text{hand}} + L_{\text{face}} + L_h + L_f \tag{6}$$

$$L_{\text{body}} = L_{\text{reproj}} + L_{\text{3D Joints}} + L_{\text{SMPL-X}} \tag{7}$$

$$L_{\text{3D Joints}} + L_{\text{SMPL-X}} = \sum_{j=1}^{J} \left\| \hat{\boldsymbol{X}}_j - \boldsymbol{X}_j \right\|_1 + \left\| \left\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}\right\} - \{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}\} \right\|_2^2 \tag{8}$$

$$L_{\text{reproj}} = \sum_{j=1}^{J} \upsilon_j \left\| \hat{\boldsymbol{x}}_j - \boldsymbol{x}_j \right\|_1. \tag{9}$$

We use υ_j as a binary variable denoting visibility of each of the J joints. The re-projection losses L_h and L_f are applied in the hand and face image coordinate space, using the affine transformations T_h, T_f. The reason for this extra penalty is that alignment errors in the 2D projection of the fingers or the facial landmarks have a much smaller magnitude compared to those of the main body joints when computed on the body image I_b

$$L_h = \sum_{j \in \text{Hand}} \upsilon_j \left\| T_h T_b^{-1} \left(\hat{\boldsymbol{x}}_j - \boldsymbol{x}_j\right) \right\|_1, \; L_f = \sum_{j \in \text{Face}} \upsilon_j \left\| T_f T_b^{-1} \left(\hat{\boldsymbol{x}}_j - \boldsymbol{x}_j\right) \right\|_1. \tag{10}$$

Fig. 3: *Left:* Example curated expressive fit. *Middle:* Hands sampled from the FreiHAND dataset [105]. *Right:* Head training data produced by running RingNet [77] on FFHQ [41] and then fitting to 2D landmarks predicted by [9].

For the hand and head only data we also employ a re-projection loss, using only the subset of joints of each part, and parameter losses:

$$L_{\text{hand}} = L_{\text{reproj}} + \left\| \left\{ \hat{\boldsymbol{\beta}}_h, \hat{\boldsymbol{\theta}}_h \right\} - \{\boldsymbol{\beta}_h, \boldsymbol{\theta}_h\} \right\|_2^2 \tag{11}$$

$$L_{\text{face}} = L_{\text{reproj}} + \left\| \left\{ \hat{\boldsymbol{\beta}}_f, \hat{\boldsymbol{\theta}}_f, \hat{\boldsymbol{\psi}}_f \right\} - \{\boldsymbol{\beta}_f, \boldsymbol{\theta}_f, \boldsymbol{\psi}_f\} \right\|_2^2 . \tag{12}$$

3.3 Implementation Details

Datasets: We curate a dataset of SMPL-X fits by running vanilla SMPLify-X [67] on the LSP [35], LSP extended [36] and MPII [3] datasets. We then ask human annotators whether the resulting body mesh is plausible and agrees with the image and collect $32,617$ pairs of images and SMPL-X parameters. To augment the training data for the body we transfer the public fits of SPIN [46] from SMPL to SMPL-X, see Sup. Mat. Moreover, we use H3.6M [31] for additional 3D supervision for the body. For the hand sub-network we employ the hand-only data of FreiHAND [105]. For the face sub-network we create a pseudo ground-truth face dataset by running RingNet [77] on FFHQ [41]. The regressed FLAME [54] parameters are refined by fitting to facial landmarks [9] for better alignment with the image and more detailed expressions. Figure 3 shows samples from all training datasets.

Architecture: For the body network we extract features $\phi \in \mathbb{R}^{2048}$ with HRNet [84]. For the face and hand sub-networks we use a ResNet18 [28] to limit the computational cost. For all networks, rather than directly regressing the parameters Θ from ϕ, we follow the iterative process of [39]. We start from an initial estimate $\Theta_0 = \bar{\Theta}$, where $\bar{\Theta}$ represents the mean, which is concatenated to the features ϕ and fed to an MLP that predicts a residual $\Delta\Theta_1 = \text{MLP}\left([\phi, \Theta_0]\right)$. The new parameter value is now equal to $\Theta_1 = \Theta_0 + \Delta\Theta_1$ and the whole process is repeated. As in [39], we iterate for $t = 3$ times. The entire pipeline is implemented in PyTorch [66]. For architecture details see Sup. Mat.

Data Pre-processing and Augmentation: We follow the pre-processing and augmentation protocol of [46] for all networks. To make the model robust to

partially visible bodies we adopt the cropping augmentation of Joo et al. [37]. In addition, we augment the hand- and face-only images with random translations, as well as down-sampling by a random factor and then up-sampling back to the original resolution. The former simulates a misaligned body prediction, while the latter bridges the gap in image quality between the full-body and part-specific data. Hand and especially face images usually have a much higher resolution and quality compared to a crop extracted from a full-body image. To simulate body conditioning for the hand- and head-only data we add random noise to the initial point of the iterative regressor. For the hands we replace the default finger pose with a random rotation r_{finger} sampled from the PCA pose space of MANO. For the head we replace the default jaw rotation $\bar{\boldsymbol{\theta}}_f$ with a random rotation of $r_{\text{f}} \sim \mathcal{U}(0, 45)$ degrees around the x-axis. For both parts, we replace their global rotation with a random rotation with angle $r_{\text{global}} \sim \mathcal{U}(r_{\min}, r_{\max})$ and the same axis of rotation as the corresponding ground-truth. We set $(r_{\min}, r_{\max})_{\text{hand}}$ to $(-90, 90)$ and $(r_{\min}, r_{\max})_{\text{face}}$ to $(-45, 45)$ degrees. The default mean shape is replaced with a random vector $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, I), I \in \mathbb{R}^{10\times 10}$ and the default neutral expression with a random expression $\boldsymbol{\psi} \sim \mathcal{N}(0, \mathcal{I})$. Some visualizations of the different types of data augmentation can be found in Sup. Mat.

Training: We first pre-train the body, hand and face networks separately, using ADAM [43], on the respective part-only datasets. We then fine-tune all networks jointly on the union of all training data, following Section 3.2, letting the network make even better use of the conditioning (see Sec. 4 and Tab. 2). Please note that for this fine-tuning, our new dataset of curated SMPL-X fits plays an instrumental role. Our exact hyper-parameters are included in the released training code.

4 Experiments

4.1 Evaluation Datasets

We evaluate on several datasets:

Expressive Hands and Faces (EHF) [67] consists of 100 RGB images paired with SMPL-X registrations to synchronized 3D scans. It contains a single subject performing a variety of interesting body poses, hand gestures and facial expressions. We use it to evaluate our whole-body predictions.

3D Poses in the Wild (3DPW) [60] consists of in-the-wild RGB video sequences annotated with 3D SMPL poses. It contains several actors performing various motions, in both indoor and outdoor environments. It is captured using a single RGB camera and IMUs mounted on the subjects. We use it to evaluate our predictions for the main body area, excluding the head and hands.

FreiHAND [105] is a multi-view RGB hand dataset that contains 3D MANO hand pose and shape annotations. The ground-truth for the test data is held-out and evaluation is performed by submitting the estimated hand meshes to an online server. We use it to evaluate our hand sub-network predictions.

Stirling/ESRC 3D [15] consists of facial RGB images with ground-truth 3D face scans. It contains 2000 neutral faces images, namely 656 high-quality

(HQ) ones and 1344 low-quality (LQ) ones. We use it to evaluate our face sub-network following the protocol of [15].

4.2 Evaluation Metrics

We employ several common metrics below. We report errors with and without rigid alignment to the ground-truth. A "PA" prefix denotes that the metric measures error after solving for rotation, scale and translation using Procrustes Alignment.

To compare with ground-truth 3D skeletons, we use the **Mean Per-Joint Position Error (MPJPE)**. For this, we first compute the 14 LSP-common joints, by applying a linear joint regressor on the ground-truth and estimated meshes, and then compute their mean Euclidean distance.

For comparing to ground-truth meshes, we use the **Vertex-to-Vertex (V2V)** error, i.e. the mean distance between the ground-truth and predicted mesh vertices. This is appropriate when the predicted and ground-truth meshes have the same topology, e.g. SMPL-X for our overall network, MANO for our hand and FLAME for our face sub-network. For a fair comparison to methods that predict SMPL instead of SMPL-X, like [39, 46], we also report V2V only on the main-body, i.e. without the hands and the head, as SMPL and SMPL-X share common topology for this subset of vertices.

For comparing to approaches that output meshes with different topology, like MTC [95] that uses the Adam model and not SMPL-X, we cannot use V2V. Instead, we compute the (mesh-to-mesh) **point-to-surface** (PtS) distance from the ground-truth mesh, as a common reference, to the estimated mesh.

For evaluation on datasets that include ground-truth scans, we compute a **scan-to-mesh** version of the above **point-to-surface** distance, namely from the ground-truth scan points to the estimated mesh surface. We use this for the face dataset of [15] to evaluate the head estimation of our face sub-network.

Finally, for the FreiHAND dataset [105] we report all metrics returned by their evaluation server. Apart from PA-MPJPE and PA-V2V described above, we also report the **F-score** [44].

4.3 Quantitative and Qualitative Experiments

First, we evaluate our approach on the 3DPW dataset that includes SMPL ground-truth meshes. Although this does not include ground-truth hands and faces, it is ideal for comparing main-body reconstruction against state-of-the-art approaches, namely HMR [39] and SPIN [46]. Table 1 presents the results, and shows that ExPose outperforms HMR and is on par with the more recent SPIN. This confirms that ExPose provides a solid foundation upon which to build detailed reconstruction for the hands and face.

We then evaluate on the EHF dataset that includes high-quality SMPL-X ground truth. This allows evaluation for the more challenging task of holistic body reconstruction, including expressive hands and face. Table 2 presents an

Table 1: Comparison on the 3DPW dataset [60] with two state-of-the-art approaches for SMPL regression, HMR [39] and SPIN [46]. The numbers are per-joint and per-vertex errors (in mm) for the body part of SMPL. ExPose outperforms HMR and is on par with SPIN, while also being able to capture details for the hands and the face.

Method	PA-MPJPE (mm)	MPJPE (mm)	PA-Body V2V (mm)
HMR [39]	81.3	130	65.2
SPIN [46]	59.2	96.9	53.0
ExPose	60.7	93.4	55.6

Table 2: Ablative study on the EHF dataset. The results are vertex-to-vertex errors expressed in mm for the different parts (i.e., all vertices, body vertices, hand vertices and head vertices). We report results for the initial body network applied on the low resolution (first row), for a version that uses the body-driven attention to estimate hands and faces (second row), and for the final regressor that jointly fine-tunes the body, hands and face sub-networks.

Networks	Attention on high-res. crops	End-to-end fine-tuning	PA-V2V (mm)			
			All	Body	L/R hand	Face
Body only	✗	✗	57.3	55.9	14.3 / 14.8	5.8
Body & Hand & Face	✓	✗	56.4	52.6	14.1 / 13.9	6.0
Body & Hand & Face	✓	✓	54.5	52.6	13.1 / 12.5	5.8

ablation study for our main components. In the first row, we see that the initial body network, that uses a low-resolution body-crop image as input, performs well for body reconstruction but makes mistakes with the hands. The next two rows add *body-driven attention*; they use the body network prediction to locate the hands and face, and then redirect the attention in the original image, crop higher-resolution image patches for them, and feed them to the respective hand and face sub-networks to refine their predictions, while initializing/conditioning their predictions. This conditioning can take place in two ways. The second row shows a naive combination using independently trained sub-networks. This fails to significantly improve the results, since there is a domain gap between images of face- or hand-only [15, 105] training datasets and hand/head image crops from full-body [3, 35, 36] training datasets; the former tend to be of higher resolution and better image quality. Please note that this is similar to [10], but extended for 3D mesh regression. In the third row, the entire pipeline is fine-tuned end-to-end. This results in a boost in quantitative performance, improving mainly hand articulation (best overall performance).

Next, we compare to state-of-the-art approaches again on the EHF dataset. First, we compare against the most relevant baseline, SMPLify-X [67], which estimates SMPL-X using an optimization approach. Second, we compare against Monocular Total Capture (MTC) [95], which estimates expressive 3D humans using the Adam model. Note that we use their publicly available implementation, which does not include an expressive face model. Third, we compare

Table 3: Comparison with the state-of-the-art approaches on the EHF dataset. The metrics are defined in Sec. 4.2. For SMPLify-X, the results reported in [67] (first row) are generated using ground truth camera parameters, so they are not directly comparable with the other approaches. MTC running time includes calculation of part orientation fields and Adam fitting. The regression based methods require extra processing to obtain input human bounding box. For example, if one uses Mask-RCNN [27] with a ResNet50-FPN [56] from Detectron2 [94] the complete running time of these methods increases by 43 milliseconds. All timings were done with a Intel Xeon W-2123 3.60GHz CPU and a Quadro P5000 GPU and are for estimating one person.

Method	Time (s)	PA-V2V (mm)				PA MPJPE (mm)		PA PtS (mm)	
		All	Body	L/R hand	Face	Body Joints	L/R hand	Mean	Median
SMPLify-X′ [67]	40-60	**52.9**	56.37	**11.4**/12.6	**5.3**	73.5	**11.9**/13.2	**28.9**	18.1
HMR [39]	0.06	N/A	67.2	N/A	N/A	82.0	N/A	34.5	21.5
SPIN [46]	0.01	N/A	60.6	N/A	N/A	102.9	N/A	40.8	28.7
SMPLify-X [67]	40-60	65.3	75.4	11.6/12.9	6.3	87.6	12.2/13.5	36.8	23.0
MTC [95]	20	67.2	N/A	N/A	N/A	107.8	16.3/17.0	41.3	29.0
ExPose (Ours)	0.16	54.5	**52.6**	13.1/**12.5**	5.8	**62.8**	13.5/**12.7**	**28.9**	**18.0**

against HMR [39] and SPIN [46], which estimate SMPL bodies, therefore we perform body-only evaluation, excluding the hand and head regions. We summarize all evaluations in Table 3. We find that ExPose outperforms the other baselines, both in terms of full expressive human reconstruction and body-only reconstruction. SMPLify-X performs a bit better locally, i.e. for the hands and face, but the full body pose can be inaccurate, mainly due to errors in OpenPose detections. In contrast, our regression-based approach is a bit less accurate locally for the hands and face, but overall it is more robust than SMPLify-X. The two approaches could be combined, with ExPose replacing the heuristic initialization of SMPLify-X with its more robust estimation; we speculate that this would improve both the accuracy and the convergence speed of SMPLify-X. Furthermore, ExPose outperforms MTC across all metrics. Finally, it is approximately two orders of magnitude faster than both SMPLify-X and MTC, which are both optimization-based approaches.

We also evaluate each sub-network on the corresponding part-only datasets. For the hands we evaluate on the FreiHAND dataset [105], and for faces on the Stirling/ESRC 3D dataset [15]. Table 4 summarizes all evaluations. The part subnetworks of ExPose match or come close to the performance of state-of-the-art methods. We expect that using a deeper backbone, e.g. a ResNet50, would be beneficial, but at a higher computational cost.

The quantitative findings of Table 2 are reflected in qualitative results. In Figure 4, we compare our final results with the initial baseline that regresses all SMPL-X parameters directly from a low-resolution image without any attention (first row in Tab. 2). We observe that our body-attention mechanism gives a clear improvement for the hand and the face area. Figure 5 contains ExPose

Table 4: We evaluate our final hand sub-network on the FreiHAND dataset [105] and the face sub-network on the test dataset of Feng et al. [15]. The final part networks are on par with existing methods, despite using a shallower backbone, i.e. a ResNet-18 vs a Resnet-50.

FreiHAND	PA-MPJPE (mm)	PA-V2V (mm)	F@5mm	F@15mm
MANO CNN [105]	11.0	10.9	0.516	0.934
ExPose hand sub-network h	12.2	11.8	0.484	0.918
Stirling3D Dataset LQ/HQ	Mean (mm)	Median (mm)	Standard Deviation (mm)	
RingNet [77]	2.08/2.02	1.63/1.58	1.79/1.69	
ExPose face sub-network f	2.27/2.42	1.76/1.91	1.97/2.03	

reconstructions, seen from multiple views, where we again see the higher level of detail offered by our method. For more qualitative results, see Sup. Mat.

5 Conclusion

In this paper, we present a regression approach for holistic expressive body reconstruction. Considering the different scale of the individual parts and the limited training data, we identify that the naive approach of regressing a holistic reconstruction from a low-resolution body image misses fine details in the hands and face. To improve our regression approach, we investigate a body-driven attention scheme. This results in consistently better reconstructions. Although the pure optimization-based approach [67] recovers the finer details, it is too slow to be practical. ExPose provides competitive results, while more than two orders of magnitude faster than [67]. Eventually the two approaches could be combined effectively, as in [46]. Considering the level of the accuracy and the speed of our approach, we believe it should be a valuable tool and enable many applications that require expressive human pose information. Future work will extend the inference to multiple humans [34, 98, 99] and video sequences [40, 45]. The rich body representation will also accelerate research on human-scene [25, 79] interaction, human-object [55, 88] interaction, and person-person interaction [16, 52]. We also plan to improve body shape estimation and the pixel-level alignment to the image.

Acknowledgements: We thank Haiwen Feng for the FLAME fits, Nikos Kolotouros, Muhammed Kocabas and Nikos Athanasiou for helpful discussions, Mason Landry and Valerie Callaghan for video voiceovers. This research was partially supported by the Max Planck ETH Center for Learning Systems. ***Disclaimer:*** MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI. MJB has financial interests in Amazon and Meshcapade GmbH.

Fig. 4: *Left:* The input image. *Middle:* Naive regression from a single body image fails to capture detailed finger articulation and facial expressions. *Right:* ExPose is able to recover these details, thanks to its attention mechanism, and produces results of similar quality as SMPLify-X, while being 200× times faster, as seen in Table 3.

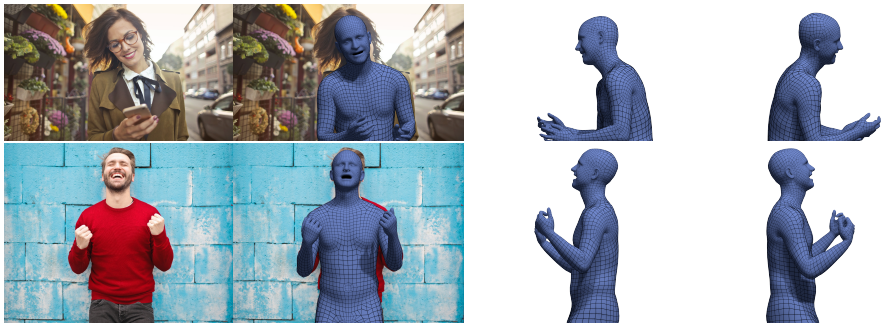

Fig. 5: Input image, ExPose predictions overlayed on the image and renderings from different viewpoints. ExPose is able to recover detailed hands and faces thanks to its attention mechanism, and produces results of similar quality as SMPLify-X, while being 200× times faster.

References

1. Agarwal, A., Triggs, B.: Recovering 3D human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **28**(1), 44–58 (2006)
2. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1446–1455 (2015)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3686–3693 (2014)
4. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: Shape Completion and Animation of PEople. ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH **24**(3), 408–416 (2005)
5. Baek, S., Kim, K.I., Kim, T.K.: Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1067–1076 (2019)
6. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: Proceedings of ACM SIGGRAPH. pp. 187–194 (1999)
7. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: European Conference on Computer Vision (ECCV). pp. 561–578 (2016)
8. Boukhayma, A., Bem, R.d., Torr, P.H.: 3D hand shape and pose from images in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10835–10844 (2019)
9. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1021–1030 (2017)
10. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2019)
11. Chandran, P., Bradley, D., Gross, M., Beeler, T.: Attention-driven cropping for very high resolution facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5861–5870 (2020)
12. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5669–5678 (2017)
13. Egger, B., Smith, W.A.P., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., Vetter, T.: 3D morphable face models–past, present and future. ACM Transactions on Graphics (TOG) **39**(5), 1–38 (2020)
14. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: A review. Computer Vision and Image Understanding (CVIU) **108**(1-2), 52–73 (2007)
15. Feng, Z.H., Huber, P., Kittler, J., Hancock, P., Wu, X.J., Zhao, Q., Koppen, P., Rätsch, M.: Evaluation of dense 3D reconstruction from 2D face images in

the wild. In: International Conference on Automatic Face & Gesture Recognition (FG). pp. 780–786 (2018)
16. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7214–7223 (2020)
17. Gabeur, V., Franco, J.S., Martin, X., Schmid, C., Rogez, G.: Moulding humans: Non-parametric 3D human shape estimation from single images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2232–2241 (2019)
18. Gavrila, D.M.: The visual analysis of human movement: A survey. Computer Vision and Image Understanding (CVIU) **73**(1), 82 – 98 (1999)
19. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., Yuan, J.: 3D hand shape and pose estimation from a single RGB image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10825–10834 (2019)
20. Grauman, K., Shakhnarovich, G., Darrell, T.: Inferring 3D structure with a statistical image-based shape model. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 641–647 (2003)
21. Guan, P., Weiss, A., Balan, A., Black, M.J.: Estimating human shape and pose from a single image. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1381–1388 (2009)
22. Guler, R.A., Kokkinos, I.: HoloPose: Holistic 3D human reconstruction in-the-wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10876–10886 (2019)
23. Güler, R.A., Neverova, N., Kokkinos, I.: DensePose: Dense human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7297–7306 (2018)
24. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: HOnnotate: A method for 3D annotation of hand and object poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3196–3206 (2020)
25. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constrains. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2282–2292 (2019)
26. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11807–11816 (2019)
27. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
29. Hidalgo, G., Raaj, Y., Idrees, H., Xiang, D., Joo, H., Simon, T., Sheikh, Y.: Single-network whole-body pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 6981–6990 (2019)
30. Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J.: Towards accurate marker-less human shape and pose estimation over time. In: International Conference on 3D Vision (3DV). pp. 421–430 (2017)

31. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **36**(7), 1325–1339 (2014)
32. Iqbal, U., Molchanov, P., Breuel, T., Gall, J., Kautz, J.: Hand pose estimation via latent 2.5D heatmap regression. In: European Conference on Computer Vision (ECCV). pp. 125–143 (2018)
33. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 2017–2025 (2015)
34. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5579–5588 (2020)
35. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 12.1–12.11 (2010)
36. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1465–1472 (2011)
37. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. arXiv preprint arXiv:2004.03686 (2020)
38. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3D deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8320–8329 (2018)
39. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131 (2018)
40. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3D human dynamics from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5607–5616 (2019)
41. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4396–4405 (2019)
42. Khamis, S., Taylor, J., Shotton, J., Keskin, C., Izadi, S., Fitzgibbon, A.: Learning an efficient model of hand shape variation from depth images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2540–2548 (2015)
43. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
44. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017)
45. Kocabas, M., Athanasiou, N., Black, M.J.: VIBE: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5253–5263 (2020)
46. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2252–2261 (2019)

47. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4496–4505 (2019)
48. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 1097–1105 (2012)
49. Kulon, D., Guler, R.A., Kokkinos, I., Bronstein, M.M., Zafeiriou, S.: Weakly-supervised mesh-convolutional hand reconstruction in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4990–5000 (2020)
50. Kulon, D., Wang, H., Güler, R.A., Bronstein, M.M., Zafeiriou, S.: Single image 3D hand reconstruction with mesh convolutions. In: Proceedings of the British Machine Vision Conference (BMVC) (2019)
51. Lee, H.J., Chen, Z.: Determination of 3D human body postures from a single view. Computer Vision, Graphics, and Image Processing **30**(2), 148–168 (1985)
52. Li, K., Mao, Y., Liu, Y., Shao, R., Liu, Y.: Full-body motion capture for multiple closely interacting persons. Graphical Models **110**, 101072 (2020)
53. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3D human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2848–2856 (2015)
54. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics (ToG) **36**(6), 194:1–194:17 (2017)
55. Li, Z., Sedlar, J., Carpentier, J., Laptev, I., Mansard, N., Sivic, J.: Estimating 3D motion and forces of person-object interactions from monocular video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8632–8641 (2019)
56. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944 (2017)
57. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755 (2014)
58. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 3730–3738 (2015)
59. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia **34**(6), 248:1–248:16 (2015)
60. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: European Conference on Computer Vision (ECCV). pp. 614–631 (2018)
61. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2659–2668 (2017)
62. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding (CVIU) **104**(2), 90–126 (2006)
63. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: GANerated hands for real-time 3D hand tracking from monocular

RGB. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 49–59 (2018)
64. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV). pp. 483–499 (2016)
65. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P.V., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: International Conference on 3D Vision (3DV). pp. 484–494 (2018)
66. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 8024–8035 (2019)
67. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10967–10977 (2019)
68. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1263–1272 (2017)
69. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3D human pose and shape from a single color image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 459–468 (2018)
70. Robinette, K.M., Blackwell, S., Daanen, H., Boehmer, M., Fleming, S., Brill, T., Hoeferlin, D., Burnsides, D.: Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Tech. Rep. AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory (2002)
71. Rogez, G., Schmid, C.: MoCap-guided data augmentation for 3D pose estimation in the wild. In: Advances in Neural Information Processing Systems (NIPS). pp. 3108–3116 (2016)
72. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia **36**(6), 245:1–245:17 (2017)
73. Rong, Y., Liu, Z., Li, C., Cao, K., Loy, C.C.: Delving deep into hybrid annotations for 3D human recovery in the wild. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5339–5347 (2019)
74. Rueegg, N., Lassner, C., Black, M.J., Schindler, K.: Chained representation cycling: Learning to estimate 3D human pose and shape by cycling between representations. In: AAAI Conference on Artificial Intelligence (AAAI) (2020)
75. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2304–2314 (2019)
76. Saito, S., Simon, T., Saragih, J., Joo, H.: PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 84–93 (2020)

77. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3D face shape and expression from an image without 3D supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7763–7772 (2019)
78. Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A.: 3D human pose estimation: A review of the literature and analysis of covariates. Computer Vision and Image Understanding (CVIU) **152**, 1–20 (2016)
79. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: PiGraphs: Learning interaction snapshots from observations. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)
80. Sigal, L., Balan, A., Black, M.J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision (IJCV) **87**(1), 4–27 (2010)
81. Sigal, L., Black, M.J.: Predicting 3D people from 2D pictures. In: International Conference on Articulated Motion and Deformable Objects. pp. 185–195 (2006)
82. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4645–4653 (2017)
83. Smith, D., Loper, M., Hu, X., Mavroidis, P., Romero, J.: FACSIMILE: Fast and accurate scans from an image in less than a second. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5329–5338 (2019)
84. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5686–5696 (2019)
85. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2621–2630 (2017)
86. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: European Conference on Computer Vision (ECCV). pp. 536–553 (2018)
87. Supančič III, J.S., Rogez, G., Yang, Y., Shotton, J., Ramanan, D.: Depth-based hand pose estimation: Data, methods, and challenges. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1868–1876 (2015)
88. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV) (2020)
89. Tekin, B., Bogo, F., Pollefeys, M.: H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4506–4515 (2019)
90. Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P.: Structured prediction of 3D human pose with deep neural networks. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 130.1–130.11 (2016)
91. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3D pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5689–5698 (2017)
92. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: BodyNet: Volumetric inference of 3D human body shapes. In: European Conference on Computer Vision (ECCV). pp. 20–38 (2018)
93. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4732 (2016)

94. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
95. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10957–10966 (2019)
96. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: Generative 3D human shape and articulated pose models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7214–7223 (2020)
97. Yuan, S., Garcia-Hernando, G., Stenger, B., Moon, G., Chang, J., Lee, K., Molchanov, P., Kautz, J., Honari, S., Ge, L., Yuan, J., Chen, X., Wang, G., Yang, F., Akiyama, K., Wu, Y., Wan, Q., Madadi, M., Escalera, S., Kim, T.K.: Depth-based 3D hand pose estimation: From current achievements to future goals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2636–2645 (2018)
98. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3D pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2148–2157 (2018)
99. Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.I., Sminchisescu, C.: Deep network for the integrated 3D sensing of multiple people in natural images. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 8410–8419 (2018)
100. Zhang, X., Li, Q., Mo, H., Zhang, W., Zheng, W.: End-to-end hand mesh recovery from a monocular RGB image. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2354–2364 (2019)
101. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3D human pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3420–3430 (2019)
102. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: DeepHuman: 3D human reconstruction from a single image. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 7738–7748 (2019)
103. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5738–5746 (2019)
104. Zimmermann, C., Brox, T.: Learning to estimate 3D hand pose from single RGB images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4913–4921 (2017)
105. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 813–822 (2019)
106. Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3D face reconstruction, tracking, and applications. Computer Graphics Forum **37**(2), 523–550 (2018)