Collaborative Regression of Expressive Bodies using Moderation

Yao Feng* Vasileios Choutas* Timo Bolkart Dimitrios Tzionas Michael J. Black
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{yfeng, vchoutas, tbolkart, dtzionas, black}@tuebingen.mpg.de
* Equal contribution

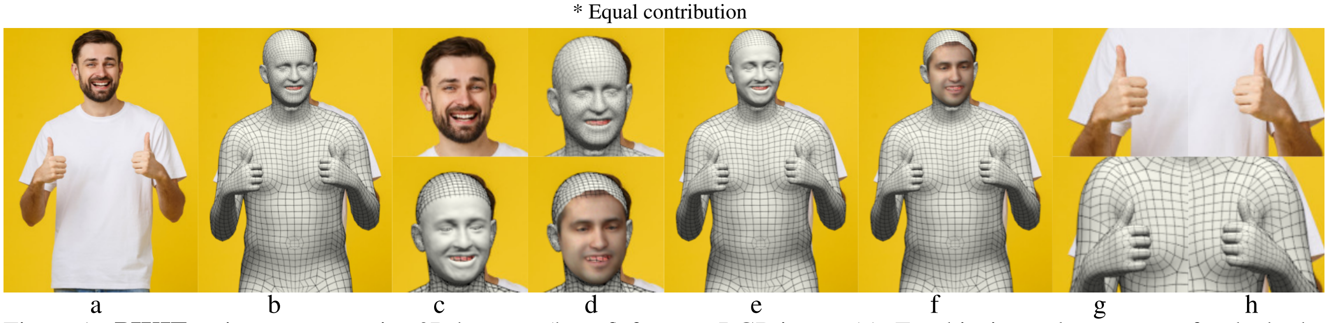

Figure 1: **PIXIE** estimates expressive 3D humans (b, e, f) from an RGB image (a). For this, it employs experts for the body, face (c, d), and hands (g, h), which are combined (b, e, f) by a novel moderator, according to their confidence (see Fig. 2). PIXIE estimates appropriate body shapes (b) by implicitly learning to reason about gender from an image. Finally, PIXIE estimates fine facial details, i.e. 3D surface displacements (c) and albedo (d), similar to state-of-the-art face-only methods.

Abstract

Recovering expressive humans from images is essential for understanding human behavior. Methods that estimate 3D bodies, faces, or hands have progressed significantly, yet separately. Face methods recover accurate 3D shape and geometric details, but need a tight crop and struggle with extreme views and low resolution. Whole-body methods are robust to a wide range of poses and resolutions, but provide only a rough 3D face shape without details like wrinkles. To get the best of both worlds, we introduce PIXIE, which produces animatable, whole-body 3D avatars with realistic facial detail, from a single image. For this, PIXIE uses two key observations. First, existing work combines independent estimates from body, face, and hand experts, by trusting them equally. PIXIE introduces a novel moderator that merges the features of the experts, weighted by their confidence. All part experts can contribute to the whole, using SMPL-X's shared shape space across all body parts. Second, human shape is highly correlated with gender, but existing work ignores this. We label training images as male, female, or non-binary, and train PIXIE to infer "gendered" 3D body shapes with a novel shape loss. In addition to 3D body pose and shape parameters, PIXIE estimates expression, illumination, albedo and 3D facial surface displacements. Quantitative and qualitative evaluation shows that PIXIE estimates more accurate whole-body shape and detailed face shape than the state of the art. Models and code are available at `pixie.is.tue.mpg.de`.

1. Introduction

To model human behavior, we need to capture how people look, how they feel, and how they interact with each other. To facilitate this, our goal is to reconstruct whole-body 3D shape and pose, facial expressions, and hand gestures from an RGB image. This is challenging, as humans vary in shape and appearance, they are highly articulated, they wear complex clothing, they are often occluded, and their face and hands are small, yet highly deformable. For these reasons, the community studies the body [13, 47, 52], hands [14, 29, 38, 110] and face [21] mostly separately.

Recent whole-body statistical models [46, 70, 104] enable approaches to address the problem holistically, by jointly capturing the body, face and hands. ExPose [18] reconstructs SMPL-X [70] meshes from an RGB image, using "expert" sub-networks for the body, face and hands. However, ExPose's part experts operate completely independently, as they only "see" their respective part image. Thus, they do not exploit the correlations between parts to overcome challenges like occlusion or motion blur.

Face-only methods [22, 105] are well studied and recover accurate facial shape, albedo and geometric details, which are important to capture emotions. However, they need a tight crop around the face and struggle with extreme viewing angles and faces that are small, low-resolution or occluded. While whole-body methods [18, 46, 70, 77, 104] handle these challenges well, they estimate average-looking

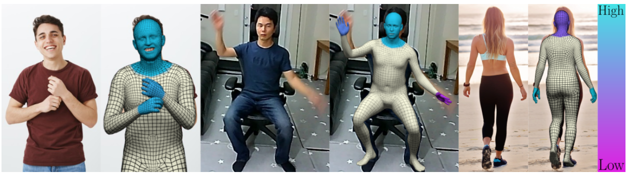

Figure 2: PIXIE infers the confidence of its body, face and hand experts, and fuses their features accordingly. Challenges, like occlusions, are resolved with full-body context. (L) Input image. (R) Color-coded part-expert confidence.

face shapes, without face albedo and fine geometric details.

To get the best of all worlds, we introduce PIXIE ("Pixels to Individuals: eXpressive Image-based Estimation"). PIXIE estimates expressive whole-bodied 3D humans from an RGB image more realistically than existing work. To do so, it pushes the state of the art in three ways.

First, PIXIE learns not only experts for the body, face and hands, but also a novel moderator that estimates their confidence in each sub-image, and fuses their features weighted by this. The learned fusion helps improve whole-body shape, using SMPL-X's shared shape space across all body parts. Moreover, it helps to robustly estimate head and hand pose when these are ambiguous (e.g. occlusions or blur) by using full-body context; see Fig. 2 for examples.

Second, PIXIE significantly improves "gendered" body shape realism. While human shape is highly correlated with gender, existing work ignores this and estimates inaccurate body shapes – often with the wrong gender or with a gender-neutral shape. An exception is SMPLify-X, but it uses an offline gender classifier and fits a gender-specific SMPL-X model. Instead, using a single unisex SMPL-X model enables end-to-end training of neural nets. PIXIE adopts this approach, and learns to implicitly reason about shape. For this, we define male, female, and non-binary body-shape priors within the SMPL-X shape space. At training time, given automatically created gender labels for input images, we train PIXIE to output plausible shape parameters for the specified gender. At inference time, PIXIE needs no gender labels, is applicable to any in-the-wild image, and supports non-binary genders. Note that this approach is general and is relevant for the broader community (face, body, whole-body). Body shape is also correlated with face shape [28, 35, 51]. Thus, we do the same "gendered" training for our face expert; this allows PIXIE to use face information to inform body shape. This training and network architecture significantly improves body shape both qualitatively and quantitatively.

Third, PIXIE's face expert additionally infers facial albedo and dense 3D facial-surface displacements. For this, we draw inspiration from Feng et al. [22], and go beyond them in three ways: (1) We use a whole-body shape space, rather than a face-only space, to capture correlations between the body and face shape. (2) We use photometric and identity losses on faces to inform whole-body shape. (3) We use the inferred geometric details only when the face expert is confident, as judged by the moderator. As shown in Fig. 1, this results in whole-body 3D humans with detailed faces that can be fully animated.

To summarize, here we make three key contributions: (1) We train a novel moderator, that infers the confidence of body-part experts and fuses their features weighted by this. This improves shape and pose inference under ambiguities. (2) We train the network to implicitly reason about gender, i.e. without gender labels at test time, with a novel "gendered" 3D shape loss that encourages likely body shapes. (3) We extend our face expert with branches that estimate facial albedo and 3D facial-surface displacements, enabling whole-body animation with a realistic face. PIXIE is a step towards automatic, accurate and realistic 3D avatar creation from a single RGB image. Models and code are available for research purposes mat `pixie.is.tue.mpg.de`.

2. Related work

Body reconstruction: For years, the community focused on the prediction of 2D or 3D landmarks for the body [17], face [15] and hands [84, 101], with a recent shift towards estimating 3D model parameters [13, 45, 47, 50, 68, 72, 89] or 3D surfaces [53, 60, 80, 81, 98]. One line of work simplifies the problem by using proxy representations like 2D joints [13, 32, 33, 40, 64, 72, 83, 95, 111], silhouettes [8, 40, 72], part labels [68, 79] or dense correspondences [76, 107]. These are then "lifted" to 3D, either as part of an energy term [13, 40, 106] or using a regressor [64, 68, 72, 95]. To overcome ambiguities, they use priors such as known limb lengths [57], joint angle limits [9], or a statistical body model [13, 40, 68, 70, 72] like SMPL [61] or SMPL-X [70]. While these approaches benefit from 2D annotations, they cannot overcome errors in the proxy features and do not fully exploit image context. The alternative is to directly regress 3D skeletons [58, 71, 87, 88, 91], statistical model parameters [18, 25, 45, 47, 48, 50, 52, 89], 3D meshes [53, 60], depth maps [27, 85], 3D voxels [98, 112] or distance fields [80, 81] from the image pixels.

Face reconstruction: Most modern monocular 3D face reconstruction methods estimate the parameters of a pre-computed statistical face model [21]. Similar to the body literature, this problem is tackled with both optimization [10, 12, 94, 99] and regression methods [23, 42, 82, 92]. Many learning-based approaches follow an analysis-by-synthesis strategy [20, 92, 93], which jointly estimates geometry, albedo, and lighting, to render a synthetic image [62, 73] that is compared with the input. Recent work [20, 22, 31] further employs face-recognition terms [16] during training to reconstruct more accurate facial geometry. Even geometric details, such as wrinkles, can be learned from large collections of in-the-wild images [22, 96]. We refer to Egger et al. [21] for a comprehensive overview. The major downsides of face-specific approaches are their need for tightly cropped face images and their inability to handle non-frontal images. The latter is mainly due to the lack of supervision; 2D landmarks may be missing or the face might not even be detected at all, in which case the photometric term is not applicable. By integrating face and body regression, PIXIE regresses head pose and shape robustly in situations where face-only methods fail and lets the face contribute to whole-body shape estimation.

Hand reconstruction: While hand pose estimation is most often performed from RGB-D data, there has been a recent shift towards the use of monocular RGB images [11, 14, 37, 38, 41, 55, 65, 90, 115]. Similar to the body, we split these into methods that predict 3D joints [41, 65, 90, 115], parameters of a statistical hand model [11, 14, 38, 55, 110], such as MANO [75], or a 3D surface [29, 54].

Whole-body reconstruction: Recent methods approach the problem of human reconstruction holistically. Some of these estimate 3D landmarks for the body, face and hands [43, 102], but not their 3D surface. This is addressed by whole-body statistical models [46, 70, 104], that jointly capture the 3D surface for the body, face and hands.

SMPLify-X [70] fits SMPL-X [70] to 2D body, hand and face keypoints [17] estimated in an image. Xiang et al. [103] estimate both 2D keypoints and a part orientation field and fit Adam [46] to these. Xu et al. [104] fit GHUM [104] to detected body-part image regions. While these methods work, they are based on optimization, consequently they are slow and do not scale up to large datasets.

Deep-learning methods [18, 77] tackle these limitations, and quickly regress SMPL-X parameters from an image. ExPose [18] uses "expert" sub-networks for the body, face and hands; the body expert estimates the body and rough part (hand/face) pose from the full-body image, while part experts refine the rough part poses using only local image information (hand/face crop). ExPose merges the output of its experts by always trusting them. Instead, we evaluate the confidence of each expert for each sub-image and fuse body/face and body/hand features weighted by this. To account for different body-part sizes, we use ExPose's body-driven attention, and multiple data sources for both part-only and whole-body supervision.

FrankMocap [77] is similar to ExPose and adds an (optional) optimization step to better align the estimated SMPL-X mesh with the image. Zhou et al. [114] train a network to regress a body-and-hands (SMPL+H) model [75] and the detailed MoFA [93] face model from an RGB image, following a body-part attention mechanism and multi-source training like ExPose. Note that SMPL+H and MoFA are disparate models, which are (offline) manually cut-and-stitched together. Instead, we use the whole-body SMPL-X model [70] that captures the shape of all body parts together, thus no stitching is required. Zhou et al. fuse only hand-body features in a "binary" fashion, while their face model is "disconnected" from the body. Instead, we fuse both face-body and hand-body features in a "fully analog" fusion, and thus our face expert can inform the whole-body shape. Zhou et al. have no face camera, and need PnP-RANSAC [26] and Procrustes to align their face to the image. Instead, we infer a face-specific camera and need no extra steps. Zhou et al. use a complicated architecture, with several modules that are trained separately, and is applicable only to whole bodies. Instead, we use no intermediate tasks to avoid possible sources of error and train our model end to end. Our full model is applicable to whole bodies, but the part experts are also (separately) applicable to part-only data.

3. Method

Here we introduce PIXIE, a novel model for reconstructing SMPL-X [70] humans with a realistic face from a single RGB image. It uses a set of expert sub-networks for body, face/head, and hand regression, and combines them in a bigger network architecture with three main novelties: (1) We use a novel moderator that assesses the confidence of part experts and fuses their features weighted by this, for robust inference under ambiguities, like strong occlusions. (2) We use a novel "gendered" shape loss, to improve body shape realism by learning to implicitly reason about gender. (3) In addition to the albedo predicted by our face expert, we employ the surface details branch of Feng et al. [22].

3.1. Expressive 3D Body Model

We use the expressive SMPL-X [70] body model, which captures whole-body pose and shape, including facial expressions and finger articulation. It is a differentiable function $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$, parameterized by shape $\boldsymbol{\beta}$, pose $\boldsymbol{\theta}$ and expression $\boldsymbol{\psi}$, that produces a 3D mesh. The shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{200}$ are coefficients of a linear shape space, learned from registered CAESAR [74] scans. This is a joint shape space for the body, face, and hands, naturally capturing their shape correlations. The expression parameters $\boldsymbol{\psi} \in \mathbb{R}^{50}$ are also coefficients of a low-dimensional linear space. The overall pose parameters $\boldsymbol{\theta}$ consist of body, jaw and hand pose vectors. Each joint rotation is encoded as a 6D vector [113], except for the jaw, which uses Euler angles, i.e. a 3D vector. We follow the notation of [47] and denote posed joints with $X(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{J \times 3}$, where $J = 55$. **Camera:** To reconstruct SMPL-X from images, we use the weak-perspective camera model with scale $s \in \mathbb{R}$ and translation $\boldsymbol{t} \in \mathbb{R}^2$. We denote the joints X and model vertices M projected on the image with $\boldsymbol{x} \in \mathbb{R}^{J \times 2}$ and $\boldsymbol{m} \in \mathbb{R}^{V \times 2}$.

3.2. PIXIE Architecture

PIXIE uses the architecture of Fig. 3, and is trained end to end. All model components are described below.

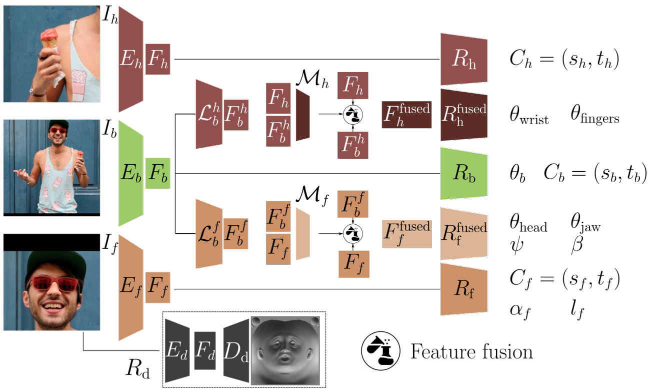

Figure 3: Body, face/head and hand image crops $\{I_b, I_f, I_h\}$ are fed to the expert encoders $\{E_b, E_f, E_h\}$ to produce part-specific features $\{F_b, F_f, F_h\}$. Our novel moderators $\{\mathcal{M}_f, \mathcal{M}_h\}$ estimate the confidence of experts for these images, and fuse face-body and hand-body features weighted by this, to create $\{F_f^{\text{fused}}, F_h^{\text{fused}}\}$. These are fed to $\{\mathcal{R}_f^{\text{fused}}, \mathcal{R}_h^{\text{fused}}\}$ for robust regression. DECA's [22] R_d estimates fine geometric details. Icon from Freepik.

Input images: Given an image I with full resolution, we assume a bounding box around the body. We use this to crop and downsample the body to I_b to feed our network. However, this makes hands and faces too low resolution. We thus use an attention mechanism [18] to extract from I high-resolution crops for the face/head, I_f, and hand, I_h.
Feature encoding: We feed $\{I_b, I_f, I_h\}$ to separate expert encoders $\{E_b, E_f, E_h\}$ to extract features $\{F_b, F_f, F_h\}$. We use ResNet-50 [39] for the face/head and hand experts to generate $F_f, F_h \in \mathcal{R}^{2048}$. The body expert E_b uses HR-Net [86], followed by convolutional layers that aggregate the multi-scale feature maps, to generate $F_b \in \mathcal{R}^{2048}$.
Feature fusion (moderator): We identify the expert pairs of {body, head} and {body, hand} as complementary, and learn the novel moderators $\{\mathcal{M}_f, \mathcal{M}_h\}$ that build "fused" features $\{F_f^{\text{fused}}, F_h^{\text{fused}}\}$ and feed them to face/head and hand regressors $\{\mathcal{R}_f^{\text{fused}}, \mathcal{R}_h^{\text{fused}}\}$ (described below) for more informed inference. A moderator is implemented as a multi-layer perceptron (MLP) and gets the body, F_b, and part, F_p (F_f or F_h), features and fuses them with a weighted sum:

$$F_p^{\text{fused}} = w_p F_b^p + (1 - w_p) F_p, \tag{1}$$

$$w_p = \frac{1}{1 + \exp\left(-t * \mathcal{M}_p(F_b^p, F_p)\right)}, \tag{2}$$

where $\mathcal{M}_p$ ($\mathcal{M}_f$ or $\mathcal{M}_h$) is the part moderator, w_p (w_f or w_h) is the expert's confidence, and F_b^p (F_b^f or F_b^h) is the body feature F_b transformed by the respective "extractor", i.e. the linear layer $\mathcal{L}^p$ ($\mathcal{L}^h$ or $\mathcal{L}^f$) between the body encoder E_b and part moderator $\mathcal{M}_p$. Finally, t is a learned temperature weight, jointly trained with all network weights with the losses of Sec. 3.3, with no t-specific supervision.

Parameter regression: We use two main regressor types: (1) We use the body, face/head, and hand $\{\mathcal{R}_b, \mathcal{R}_f, \mathcal{R}_h\}$ regressors, that get features *only* from the respective expert encoder $\{F_b, F_f, F_h\}$. $\mathcal{R}_b$ infers the camera $C_b = (s_b, \boldsymbol{t}_b)$, and body rotation and pose $\boldsymbol{\theta}_b$ up to (excluding) the head and wrist. $\mathcal{R}_f$ infers the camera $C_f = (s_f, \boldsymbol{t}_f)$, face albedo $\boldsymbol{\alpha}_f$, and lighting $\mathbf{l}_f$. $\mathcal{R}_h$ infers the camera $C_h = (s_h, \boldsymbol{t}_h)$. (2) We use the face/head, $\mathcal{R}_f^{\text{fused}}$, and hand, $\mathcal{R}_h^{\text{fused}}$, regressors that get from moderators the "fused" features, F_f^{fused} and F_h^{fused}. $\mathcal{R}_h^{\text{fused}}$ infers the wrist $\boldsymbol{\theta}_{\text{wrist}}$ and finger pose $\boldsymbol{\theta}_{\text{fingers}}$. $\mathcal{R}_f^{\text{fused}}$ infers expressions $\boldsymbol{\psi}$, head rotation $\boldsymbol{\theta}_{\text{head}}$, and jaw pose $\boldsymbol{\theta}_{\text{jaw}}$. Importantly, $\mathcal{R}_f^{\text{fused}}$ also infers body shape $\boldsymbol{\beta}$, letting our face expert contribute to whole-body shape.
Detail capture: We use the fine geometric details branch R_d of Feng et al. [22] that, given a face image I_f, estimates dense 3D displacements on top of FLAME's [59] surface. We convert the displacements from FLAME's to SMPL-X's UV map, and apply them on PIXIE's inferred head shape. However, inferring geometric details from full-body images is not trivial; faces tend to be much noisier in these compared to face-only images. We account for this with our moderator, and use the inferred displacements only when the face/head expert is confident.

3.3. Training Losses

To train PIXIE we use body, hand and face losses:

$$L = L_{\text{body}} + L_{\text{hand}} + L_{\text{face}} + L_{\text{update}}, \tag{3}$$

defined as follows; the hat (e.g. $\hat{\boldsymbol{x}}$) denotes ground truth.
Body losses: Following [18], we use a combination of a 2D re-projection, a 3D joint, and a SMPL-X parameter loss:

$$L_{\text{body}} = L_{\text{2D/3D-Joints}}^{\text{body}} + L_{\text{params}}^{\text{body}}, \tag{4}$$

$$L_{\text{2D/3D-Joints}}^{\text{body}} = \sum_{j=1}^{J} \left\|\hat{\boldsymbol{x}}_j - \boldsymbol{x}_j\right\|_1 + \sum_{j=1}^{J} \left\|\hat{\boldsymbol{X}}_j - \boldsymbol{X}_j\right\|_1, \tag{5}$$

$$L_{\text{params}}^{\text{body}} = \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2^2 + \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2. \tag{6}$$

Hand losses: We employ a similar set of losses to train the 3D hand pose and shape estimation network:

$$L_{\text{hand}} = L_{\text{2D/3D-Joints}}^{\text{hand}} + L_{\text{params}}^{\text{hand}}, \tag{7}$$

defined similarly to $L_{\text{2D/3D-Joints}}^{\text{body}}$ and $L_{\text{params}}^{\text{body}}$ of the body, but using the hand joints and pose parameters $\boldsymbol{\theta}_{\text{wrist}}$ and $\boldsymbol{\theta}_{\text{fingers}}$.
Face losses: We adopt standard losses used by the 3D face estimation community [20, 22]:

$$L_{\text{face}} = L_{\text{lmk}} + L_{\text{lmk-closure}} + L_{\text{params}}^{\text{face}} + L_{\text{pho}} + L_{\text{id}}. \tag{8}$$

The landmark loss penalizes the difference between detected [15] target 2D landmarks $\hat{\boldsymbol{m}}_j$ and respective model landmarks (lying on M_f) projected on the image plane, $\boldsymbol{m}_j$:

$$L_{\text{lmk}} = \sum_{j=1}^{N_{\text{lmk}}} \left\| \hat{\boldsymbol{m}}_j - \boldsymbol{m}_j \right\|_1 . \quad (9)$$

Following [22], we also compute a loss for the set E of landmarks on the upper, lower eyelid and upper, lower lip:

$$L_{\text{lmk-closure}} = \sum_{(i,j)\in E} \left\| (\hat{\boldsymbol{m}}_i - \hat{\boldsymbol{m}}_j) - (\boldsymbol{m}_i - \boldsymbol{m}_j) \right\|_1 . \quad (10)$$

The face parameter loss $L^{\text{face}}_{\text{params}}$ follows $L^{\text{body}}_{\text{params}}$, but for face pose $\boldsymbol{\theta}_{\text{face}}$ only. This loss is only used for face crops from body data, when the target face pose is available.

Given the predicted 3D face mesh M_f as a subset of M, face albedo $\boldsymbol{\alpha}_f$ and lighting $\mathbf{l}_f$, we render a synthetic image I_r for the input subject using the differentiable renderer from Pytorch3D [73]. We then minimize the difference between the input face image I_f and the rendered image I_r:

$$L_{\text{pho}} = \left\| \mathrm{S} \odot (I_f - I_r) \right\|_{1,1} , \quad (11)$$

where S is a binary face mask with value 1 in the face skin region, and 0 elsewhere, and $\odot$ denotes the Hadamard product. The segmentation mask prevents errors from non-face regions influencing the optimization, and we use the segmentation network of Nirkin et al. [67] to extract S. The image formation process is the same as in Feng et al. [22].

Following [20, 30], we use a pre-trained face recognition network [16], f_{id}, to compute embeddings for the rendered image I_r and the input I_f. We then maximize the cosine similarity between the two identity embeddings

$$L_{\text{id}} = 1 - \frac{< f_{\text{id}}(I_f), f_{\text{id}}(I_r) >}{\left\| f_{\text{id}}(I_f) \right\|_2 \cdot \left\| f_{\text{id}}(I_r) \right\|_2} . \quad (12)$$

Priors: Due to the difficulty of the problem, we use additional priors to constrain PIXIE to generate plausible solutions. For expression parameters, we use a Gaussian prior:

$$L_{\text{exp}}(\boldsymbol{\psi}) = \left\| \boldsymbol{\psi} \right\|_2^2 . \quad (13)$$

We also add soft regularization on jaw and face pose:

$$L_{\text{jaw}}(\boldsymbol{\theta}_{\text{jaw}}) = \left| \boldsymbol{\theta}^{\text{pitch}}_{\text{jaw}} \right|^2 + \left| \boldsymbol{\theta}^{\text{roll}}_{\text{jaw}} \right|^2 + \left| \min(\boldsymbol{\theta}^{\text{yaw}}_{\text{jaw}}, 0) \right|^2 , \quad (14)$$

$$L_{\text{face}}(\boldsymbol{\theta}_{\text{face}}) = \left| \max(\left| \boldsymbol{\theta}^{\text{yaw}}_{\text{face}} \right|, 90) \right|^2 . \quad (15)$$

All these priors are "standard" regularizers, empirically found to discourage implausible configurations (extreme values, unrealistic shape/pose, inter-penetrations, etc).

Gender: As gender strongly affects body shape, we use a gender-specific shape prior during training, when gender labels are available. For this, we register SMPL-X to CAESAR [74] scans, and compute the mean $\boldsymbol{\mu}$ and covariance Σ of shape parameters for each gender. We then use:

$$L_{\text{shape}}(\boldsymbol{\beta}) = \begin{cases} (\boldsymbol{\beta} - \mu_{\text{F}})^T \Sigma_{\text{F}} (\boldsymbol{\beta} - \mu_{\text{F}}) & \text{if female} \\ (\boldsymbol{\beta} - \mu_{\text{M}})^T \Sigma_{\text{M}} (\boldsymbol{\beta} - \mu_{\text{M}}) & \text{if male} \\ \left\| \boldsymbol{\beta} \right\|_2^2 & \text{o/w.} \end{cases} \quad (16)$$

When gender is unknown, we use a Gaussian prior computed over all scans/registrations, irrespective of gender. Please note that we do not need gender labels for inference.

Feature update loss: We encourage the transformed body features F^p_b (F^f_b or F^h_b) to match F^{fused}_p with a loss that was empirically found to stabilize network training:

$$L_{\text{update}} = \left\| F^p_b - F^{\text{fused}}_p \right\|_1 . \quad (17)$$

3.4. Implementation Details

Training data: For whole-body data we use the curated SMPL-X fits of [18], and SMPL-X fits to whole-body COCO data [43]. For hand-only data we use FreiHAND [116] and Total Motion [103]. For face/head data we use VGGFace2 [16] and detect $N_{\text{lmk}} = 68$ 2D landmarks with the method of Bulat et al. [15]. We get gender annotations by running the method of Rothe et al. [78] on many photos per identity and using majority voting to improve robustness. For data augmentation, see Sup. Mat.

Network training: We do multi-step training that empirically aids stability. We pre-train on part-only data, and train on whole-body data end to end; for details see Sup. Mat.

4. Experiments

4.1. Evaluation Datasets

EHF [70]: We evaluate whole-body accuracy on this. It has 100 RGB images of 1 minimally-clothed subject in a lab setting with ground-truth SMPL-X meshes and 3D scans.

AGORA [69]: We evaluate whole-body and body-only accuracy on this, using its body-face-hands (BFH) subset. It has rendered [6] photo-realistic images of 3D human scans [1, 2, 4, 5] in scenes [3, 7]. It has SMPL-X ground truth recovered from scans, images and semantic labels [108].

3DPW [100]: We evaluate main-body accuracy on this. It captures 5 subjects in indoor/outdoor videos with SMPL pseudo ground truth, recovered from images and IMUs.

NoW [82]: We use it to evaluate face/head-only accuracy. It contains 3D head scans for 100 subjects, and 2054 images with various viewing angles and facial expressions.

FreiHAND [116]: We evaluate hand-only accuracy on this. It has $37k$ hand/hand-object images of 32 subjects, with MANO ground truth, recovered from multi-view images.

4.2. Evaluation Metrics

Mesh alignment: Prior to computing a metric, we align estimated meshes to ground-truth ones. The prefix "PA"

Method	Type	Body model	Time (s)	PA-V2V (mm) ↓				TR-V2V (mm) ↓				PA-MPJPE (mm) ↓		PA-P2S (mm) ↓	
				All	Body	L/R hand	Face	All	Body	L/R hand	Face	MPJPE-14	L/R hand	Mean	Median
SMPLify-X′ [70]	O	SMPL-X	40-60	52.9	56.37	11.4/12.6	4.4	79.5	92.3	21.3/22.1	10.9	73.5	12.9/13.2	28.9	18.1
SMPLify-X [70]	O	SMPL-X	40-60	65.3	75.4	11.6/12.9	4.9	93.0	116.1	**23.8/24.9**	**11.5**	87.6	12.2/13.5	36.8	23.0
MTC [103]	O	Adam	20	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	107.8	16.3/17.0	41.3	29.0
SPIN [52]	R	SMPL	**0.01**	N/A	60.6	N/A	N/A	N/A	96.8	N/A	N/A	102.9	N/A	40.8	28.7
FrankMocap [77]	R	SMPL-X	0.08	57.5	52.7	12.8/12.4	N/A	76.9	80.1	32.1 / 31.9	N/A	62.3	13.2/12.6	31.6	19.2
ExPose [18]	R	SMPL-X	0.16	**54.5**	**52.6**	13.1/12.5	4.8	**65.7**	76.8	31.2 / 32.4	15.9	62.8	13.5/12.7	**28.9**	**18.0**
PIXIE (ours)	R	SMPL-X	0.08-0.10	55.0	53.0	**11.2/11.0**	**4.6**	67.6	**75.8**	25.6/27.0	14.2	**61.5**	**11.7/11.4**	29.9	18.4

Table 1: Evaluation on EHF [70]. PIXIE is on par with the state of the art w.r.t. body and face performance, but predicts better hand poses. SMPLify-X′ uses the ground-truth focal length (*excluded from bold*). Run-times were measured on an Intel Xeon W-2123 3.60GHz machine with a NVIDIA Quadro P5000 GPU. "O/R" denotes Optimization/Regression.

denotes Procrustes Alignment (solving for scale, rotation and translation), while "TR" denotes translation alignment. "TR" is stricter, as it does not factor out scale and rotation. When reporting hand-/face-only metrics for the full body, we align each part separately.

Mean Per-Joint Position Error (MPJPE): We report the mean Euclidean distance between the estimated and ground-truth joints. For the body-only metric, we compute the 14 LSP-common joints [44] as a common skeleton across different body models, using a linear joint regressor [13, 56] on the estimated and ground-truth vertices. This is a standard metric, but is too sparse; it cannot capture errors in full 3D shape (i.e. surface), or all limb rotation errors.

Vertex-to-Vertex (V2V): For methods that infer meshes with the same topology as the ground-truth ones, e.g. SMPL(-X) estimations and SMPL(-X) ground truth, we compute the mean per-vertex error by taking into account *all* vertices. This is not possible for methods with different topology, e.g. SMPL estimations for SMPL-X ground truth, and vice versa. For such cases, we compute a *main-body* variant of V2V, i.e. without the hands and head, as SMPL and SMPL-X share the same topology for the main body. FB-V2V is the weighted sum of body (B), hand (LH, RH) and face (F) errors: $\mathrm{FB} = \mathrm{B} + \frac{\mathrm{LH}+\mathrm{RH}+\mathrm{F}}{3}$. V2V is stricter than MPJPE; it also captures 3D shape errors and unnatural limb rotations (for the same joint positions).

Point-to-Surface (P2S): To compare PIXIE with methods that use a different mesh topology to SMPL(-X), e.g. MTC [103], we measure the mean distance from ground-truth vertices to the *surface* of the estimated mesh. P2S is stricter than MPJPE; it captures errors in 3D shape, but not unnatural limb rotations (for the same joint positions).

4.3. Quantitative Evaluation

Whole-body. In Tab. 1 - 2 we report whole-body metrics ("All"), by taking into account the body, face and hands jointly. We add body-only ("Body"), hand-only ("L/R hand"), and face-only ("Face") variants for completeness.

EHF [70]: Table 1 compares PIXIE to three baseline sets: (1) the optimization-based SMPLify-X [70] and MTC [103] that infer SMPL-X and Adam, (2) the regression-based SPIN [52] that infers SMPL, and (3) the regression-based ExPose [18] and FrankMocap [77] that infer SMPL-X. Note that MTC does not estimate the face. PIXIE outperforms optimization methods on most metrics, while being significantly faster. Moreover, it is on par with regression methods, both in terms of error metrics and runtime, which drops to 0.08 sec for known body-part crops.

Method	PA-V2V (mm) ↓		TR-V2V (mm) ↓	
	All	Body	All	Body
Naive Body	59.7	54.3	70.5	83.4
"Copy-paste"	60.3	55.5	72.9	82.4
PIXIE (ours)	**55.0**	**53.0**	**67.6**	**75.8**

Table 2: Ablation for our moderator on EHF [70]. "Naive body" denotes a single regressor for the whole body, and "Copy-Paste" denotes a naive integration of the independent expert estimations on the inferred body.

AGORA [69]: Figure 4 compares PIXIE to whole-body [18, 70, 77] and body-only [45, 47, 50, 52, 60, 89] regressors, for a varying occlusion degree. PIXIE outperforms all methods, and is competitive on body-only metrics even to the occlusion-aware PARE [50]. Note that AGORA is much more complex and natural than EHF, making the results more representative of real-world scenarios.

Ablation for moderators: Table 2 compares PIXIE to naive whole-body regression (no body-part experts) and the "copy-paste" fusion strategy. The latter copies pose parameters from the part experts (see [18, 77]), as well as shape parameters from the face expert, to the whole body.

The naive version does not benefit from the expertise of the part experts. "Copy-paste" fusion can lead to erroneous hand/face orientation inference, since the respective experts lack global context. Moreover, estimating whole-body shape from a face image is not always reliable, e.g. when a person faces away from the camera (Fig. 2). PIXIE fuses "global" body and "local" part features with its moderators. In this way, it estimates more accurate 3D bodies and is more robust to challenging ambiguities (blur, occlusion) than existing whole-body regressors, especially on stricter metrics without Procrustes alignment.

Ablation for "gendered" shape loss on 3DPW [100]: By

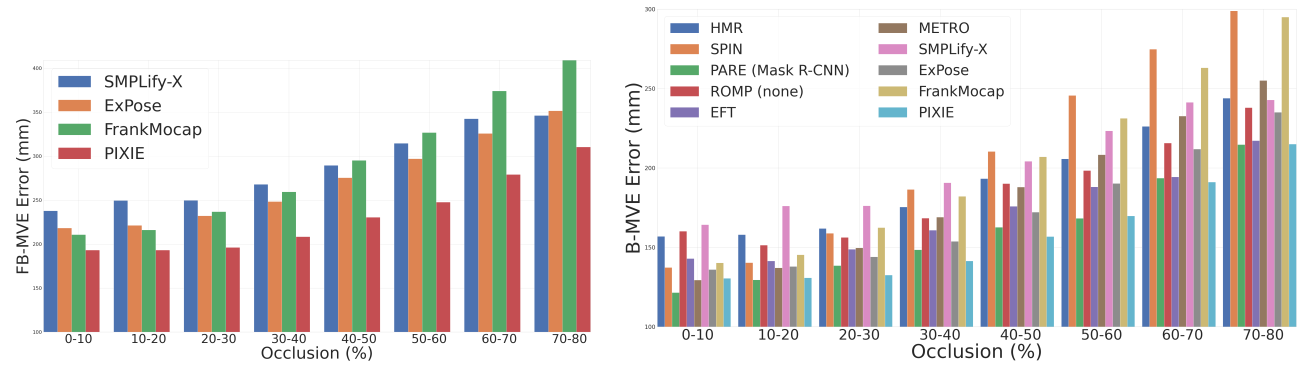

Figure 4: Comparison against state-of-the-art full-body (left) and body-only (right) methods on AGORA [69], using the vertex-to-vertex (V2V) metric (mm) for varying percentages of occlusion. Unless otherwise noted (in parens), we use OpenPose to extract person bounding boxes. PIXIE outperforms existing methods, including the occlusion-aware PARE [50].

Method	Body model	PA-MPJPE (mm) ↓	TR-MPJPE (mm) ↓	Body PA-V2V (mm) ↓
HMR [47]	SMPL	81.3	130.0	65.2
SPIN [52]	SMPL	**59.2**	96.9	53.0
FrankMocap [77]	SMPL-X	61.9	96.7	55.1
ExPose [18]	SMPL-X	60.7	93.4	55.6
PIXIE (ours)	SMPL-X	61.3	**91.0**	**50.9**

Table 3: Evaluation on 3DPW [100]. PIXIE is the best for the stricter TR-MPJPE (joints) and V2V (surface) metrics.

removing our "gendered" shape loss, the PA-V2V error increases from 50.9 to 51.7 mm. A qualitative ablation is shown in Fig. 5; learned implicit reasoning about gender gives more realistic body shapes. SMPL-X's shared shape space for the whole body lets parts contribute to the whole.

Parts-only: For completeness, we use standard benchmarks for body-only, face-only, and hand-only evaluation.

Body-only on 3DPW [100]: Table 3 shows that PIXIE performs on par FrankMocap [77] and ExPose [18] and is worse than SPIN [52], for the PA-MPJPE metric, but outperforms them all in the stricter TR-MPJPE (joints) and V2V (surface) metrics.

Face-only on NoW [82]: Table 4 shows that PIXIE outperforms not only the expressive whole-body method ExPose [18], but also strong and dedicated face-only methods, except for the recent work of Feng et al. [22].

Hand-only on FreiHAND [116]: Table 5 shows that our hand expert performs on par with the whole-body ExPose [18], is a bit worse than the hand-specific "MANO CNN" [116], but outperforms the hand expert of Zhou et al. [114].

4.4. Qualitative Evaluation

Figure 6 compares PIXIE with FrankMocap [77] and ExPose [18], which also regresses SMPL-X. Both baselines fail when the hand expert faces ambiguities (row 2); PIXIE gains robustness by using the full-body context. Both baselines give body shapes that look average (rows 1, 4) or have the wrong gender (rows 2, 3); PIXIE gives the most realistic shapes due to its "gendered" shape loss. FrankMocap fails for strong occlusions (rows 1, 3). Lastly, ExPose struggles with accurate facial expressions, and FrankMocap with head rotations (rows 1, 3); PIXIE outperforms both with its strong face/head expert and predicts a more realistic face.

Figure 7 compares PIXIE with Zhou et al. [114], recent work that also estimates a textured face. PIXIE gives more accurate poses (see how hands and faces align to the image), as it fuses both face-body and hand-body expert features, weighted by their confidence. PIXIE also gives more realistic body shapes, both due to its gendered shape loss and due to part experts contributing to whole-body shape, using SMPL-X's shared body, hand and face shape space.

Method	PA-P2S for face/head (mm) ↓		
	Median (mm) ↓	Mean (mm) ↓	Std (mm) ↓
3DMM-CNN [97]	1.84	2.33	2.05
PRNet [23]	1.50	1.98	1.88
Deng et al. [20]	1.23	1.54	1.29
RingNet [82]	1.21	1.54	1.31
3DDFA-V2 [36]	1.23	1.57	1.39
DECA [22]	**1.09**	**1.38**	**1.18**
ExPose [18]	1.26	1.57	1.32
PIXIE (ours)	*1.18*	*1.49*	*1.25*

Table 4: Evaluation on NoW [82]. PIXIE is better than the whole-body ExPose, it outperforms many strong face-specific methods, and is a bit worse than DECA [22].

Method	PA-MPJPE (mm) ↓	PA-V2V (mm) ↓	PA-F@ 5mm ↑	PA-F@ 15mm ↑
"MANO CNN" [116]	**11.0**	**10.9**	**0.516**	**0.934**
ExPose [18] hand expert	12.2	*11.8*	*0.484*	0.918
Zhou et al. [114]	15.7	-	-	-
PIXIE hand expert	*12.0*	12.1	0.468	*0.919*

Table 5: Evaluation on FreiHAND [116]. PIXIE's hand expert is on par with the hand expert of ExPose, but clearly outperforms the more related Zhou et al. [114] that also uses hand-body feature fusion.

Figure 5: Ablation for the "gendered" shape loss and the shared shape space (body/head). From left to right: (1) RGB Image, (2) shape prediction only from the body image, and PIXIE without (3) and with (4) the "gendered" shape loss. We always use the gender-neutral SMPL-X model.

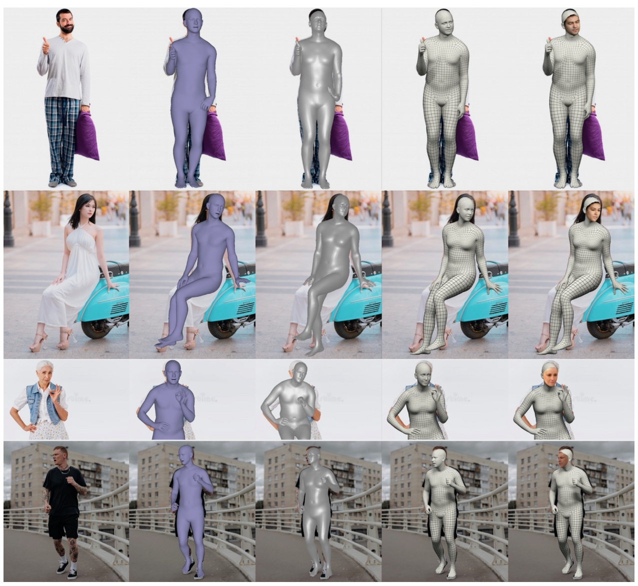

Figure 6: Qualitative comparison. From left to right: (1) RGB Image, (2) ExPose [18], (3) FrankMocap [77], (4) PIXIE, (5) PIXIE with predicted albedo and lighting.

Future work: Mesh-to-image misalignment is a common limitation of regressors that pool "global" features from the image, losing local information. This could be tackled with "pixel-aligned" features [34, 50, 80, 109]. Moreover, SMPL-X models bodies without clothing; adding clothing models [19, 63] is a challenging but promising avenue. Furthermore, due to the formulation of the photometric term the model prefers to explain image evidence using lighting, rather than albedo, which leads to wrong skin tone predictions. Future work could further improve cases with self-contact [24, 66], or other extreme ambiguities.

Figure 7: Comparison with Zhou et al. [114]. From left to right: (1) RGB image, (2) Zhou et al., (3) PIXIE with inferred facial details and (4) inferred albedo and lighting. Note that Zhou et al. use tight face crops through Dlib [49] to improve performance; PIXIE needs no tight face crops.

5. Conclusion

We present PIXIE, a novel expressive whole-body reconstruction method that recovers an animatable 3D avatar with a detailed face from a single RGB image. PIXIE uses body-driven attention to leverage dedicated body, head and face experts. It learns a novel moderator that reasons about the confidence of each expert, to fuse their features according to confidence, and exploit their complementary strengths. It uses the best practices from the face community for accurate faces with realistic albedo and geometric details. The face expert can contribute to more realistic whole-body shapes, by using a shared face-body shape space. To further improve shape, PIXIE uses implicit reasoning about gender, to encourage likely "gendered" body shapes. Qualitative results show natural and expressive humans, with improved body shape, well articulated hands, and realistic faces, comparable to the best face-only methods. We believe that PIXIE will be useful for many applications that need expressive human understanding from images.

Acknowledgments: We thank Victoria Fernández Abrevaya, Yinghao Huang, Yuliang Xiu, Radek Danecek for discussions and Priyanka Patel for AGORA experiments. This work was partially supported by the Max Planck ETH Center for Learning Systems.
Disclosure: https://files.is.tue.mpg.de/black/CoI/3DV2021.txt

References

[1] 3DPeople. 3dpeople.com. 5

[2] AXYZ. secure.axyz-design.com. 5

[3] HDRI Haven. hdrihaven.com. 5

[4] HumanAlloy. humanalloy.com. 5

[5] RenderPeople. renderpeople.com. 5

[6] Unreal Engine. unrealengine.com. 5

[7] Unreal Engine Marketplace. unrealengine.com/marketplace/en-US/store. 5

[8] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(1):44–58, 2006. 2

[9] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, 2015. 2

[10] Oswald Aldrian and William AP Smith. Inverse rendering of faces with a 3D morphable model. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(5):1080–1093, 2013. 2

[11] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1067–1076, 2019. 3

[12] Anil Bas, William A. P. Smith, Timo Bolkart, and Stefanie Wuhrer. Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision Workshops (ACCVw)*, pages 377–391, 2017. 2

[13] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, pages 561–578, 2016. 1, 2, 6

[14] Adnane Boukhayma, Rodrigo de Bem, and Philip H.S. Torr. 3D hand shape and pose from images in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10843–10852, 2019. 1, 3

[15] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017. 2, 4, 5

[16] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 67–74, 2018. 2, 5

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

[18] 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 (ECCV)*, pages 20–40, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[19] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[20] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*, pages 285–295, 2019. 2, 4, 5, 7

[21] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D morphable face models - past, present and future. *Transactions on Graphics (TOG)*, 39(5):157:1–157:38, 2020. 1, 2, 3

[22] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH*, 40(4):88:1–88:13, Aug. 2021. 1, 2, 3, 4, 5, 7

[23] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision (ECCV)*, pages 557–574, 2018. 2, 7

[24] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3D human Self-Contact. In *AAAI Conference on Artificial Intelligence*, 2021. 8

[25] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7212–7221, 2020. 2

[26] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM (CACM)*, 24(6):381–395, 1981. 3

[27] Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding Humans: Non-parametric 3D human shape estimation from single images. In *International Conference on Computer Vision (ICCV)*, pages 2232–2241, 2019. 2

[28] Alessio Gallucci, Dmitry Znamenskiy, and Milan Petkovic. Prediction of 3D body parts from face shape and anthropometric measurements. *Journal of Image and Graphics*, 8(3), 2020. 2

[29] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3D hand shape and pose estimation from a single RGB image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10833–10842, 2019. 1, 3

[30] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, 2019. 5

[31] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron

Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3D morphable model regression. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8377–8386, 2018. 2

[32] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3D structure with a statistical image-based shape model. In *International Conference on Computer Vision (ICCV)*, pages 641–648, 2003. 2

[33] Peng Guan, Alexander Weiss, Alexandru Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *International Conference on Computer Vision (ICCV)*, pages 1381–1388, 2009. 2

[34] Riza Alp Güler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction In-The-Wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10876–10886, 2019. 8

[35] Semih Gunel, Helge Rhodin, and Pascal Fua. What face and body shapes can tell us about height. In *International Conference on Computer Vision Workshops (ICCVw)*, pages 1819–1827, 2019. 2

[36] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3D dense face alignment. In *European Conference on Computer Vision (ECCV)*, volume 12364, pages 152–168, 2020. 7

[37] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2020. 3

[38] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 1, 3

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[40] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *International Conference on 3D Vision (3DV)*, pages 421–430, 2017. 2

[41] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5D heatmap regression. In *European Conference on Computer Vision (ECCV)*, pages 125–143, 2018. 3

[42] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *International Conference on Computer Vision (ICCV)*, pages 1031–1039, 2017. 2

[43] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision (ECCV)*, pages 196–214, 2020. 3, 5

[44] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*, pages 12.1–12.11, 2010. 6

[45] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar Fine-Tuning for 3D Human Pose Fitting Towards In-the-Wild 3D Human Pose Estimation. In *International Conference on 3D Vision (3DV)*, 2021. 2, 6

[46] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 1, 3

[47] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 1, 2, 3, 6, 7

[48] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5614–5623, 2019. 2

[49] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research (JMLR)*, 10:1755–1758, 2009. 8

[50] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part Attention Regressor for 3D Human Body Estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, Oct. 2021. 2, 6, 7, 8

[51] Enes Kocabey, Mustafa Camurcu, Ferda Ofli, Yusuf Aytar, Javier Marín, Antonio Torralba, and Ingmar Weber. Face-to-BMI: Using computer vision to infer body mass index on social media. In *International Conference on Web and Social Media (ICWSM)*, pages 572–575, 2017. 2

[52] 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 (ICCV)*, pages 2252–2261, 2019. 1, 2, 6, 7

[53] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4501–4510, 2019. 2

[54] Dominik Kulon, Riza Alp Güler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4989–4999, 2020. 3

[55] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael M. Bronstein, and Stefanos Zafeiriou. Single image 3D hand reconstruction with mesh convolutions. In *British Machine Vision Conference (BMVC)*, 2019. 3

[56] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the People: Closing the Loop Between 3D and 2D Human Representations. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4704–4713, 2017. 6

[57] Hsi-Jian Lee and Zen Chen. Determination of 3D human body postures from a single view. *Computer Vision, Graphics, and Image Processing*, 30(2):148–168, 1985. 2

[58] Sijin Li, Weichen Zhang, and Antoni B Chan. Maximum-margin structured learning with deep networks for 3D human pose estimation. In *International Conference on Com-

puter Vision (ICCV)*, pages 2848–2856, 2015. 2

[59] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics (TOG)*, 36(6):194:1–194:17, 2017. 4

[60] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-End Human Pose and Mesh Reconstruction with Transformers. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6

[61] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 2

[62] Matthew M. Loper and Michael J. Black. OpenDR: An approximate differentiable renderer. In *European Conference on Computer Vision (ECCV)*, pages 154–169, 2014. 2

[63] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6468–6477, 2020. 8

[64] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3D human pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017. 2

[65] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. GANerated hands for real-time 3D hand tracking from monocular RGB. In *Computer Vision and Pattern Recognition (CVPR)*, pages 49–59, 2018. 3

[66] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self contact and human pose. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, June 2021. 8

[67] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 98–105, 2018. 5

[68] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, pages 484–494, 2018. 2

[69] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 5, 6, 7

[70] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 2, 3, 5, 6

[71] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2017. 2

[72] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018. 2

[73] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv:2007.08501*, 2020. 2, 5

[74] Kathleen M. Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoeferlin, and Dennis Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Technical Report AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory, 2002. 3, 5

[75] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017. 3

[76] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3D human recovery in the wild. In *International Conference on Computer Vision (ICCV)*, pages 5339–5347, 2019. 2

[77] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In *International Conference on Computer Vision Workshops (ICCVw)*, 2021. 1, 3, 6, 7, 8

[78] Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep expectation of apparent age from a single image. In *International Conference on Computer Vision Workshops (ICCVw)*, pages 252–257, 2015. 5

[79] Nadine Rueegg, Christoph Lassner, Michael J. Black, and Konrad Schindler. Chained representation cycling: Learning to estimate 3D human pose and shape by cycling between representations. In *AAAI Conference on Artificial Intelligence*, pages 5561–5569, 2020. 2

[80] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 2, 8

[81] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 81–90, 2020. 2

[82] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, 2019. 2, 5, 7

[83] Leonid Sigal and Michael J Black. Predicting 3D people from 2D pictures. In *International Conference on Articulated Motion and Deformable Objects (AMDO)*, pages 185–195, 2006. 2

[84] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser

Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, 2017. 2

[85] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. FACSIMILE: Fast and accurate scans from an image in less than a second. In *International Conference on Computer Vision (ICCV)*, pages 5329–5338, 2019. 2

[86] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019. 4

[87] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *International Conference on Computer Vision (ICCV)*, pages 2621–2630, 2017. 2

[88] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *European Conference on Computer Vision (ECCV)*, pages 536–553, 2018. 2

[89] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei. Monocular, One-stage, regression of multiple 3D people. In *International Conference on Computer Vision (ICCV)*, pages 11179–11188, Oct. 2021. 2, 6

[90] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, 2019. 3

[91] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3D human pose with deep neural networks. In *British Machine Vision Conference (BMVC)*, 2016. 2

[92] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2549–2559, 2018. 2

[93] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. MoFA: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 3735–3744, 2017. 2, 3

[94] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 2

[95] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5689–5698, 2017. 2

[96] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3D face morphable model. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1126–1135, 2019. 2

[97] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing Robust and Discriminative 3D Morphable Models with a Very Deep Neural Network. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1502, 2017. 7

[98] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision (ECCV)*, pages 20–38, 2018. 2

[99] Thomas Vetter and Volker Blanz. Estimating coloured 3D face models from single images: An example based approach. In *European Conference on Computer Vision (ECCV)*, pages 499–513, 1998. 2

[100] Timo von Marcard, Roberto Henschel, Michael J. 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 (ECCV)*, pages 614–631, 2018. 5, 6, 7

[101] Yangang Wang, Cong Peng, and Yebin Liu. Mask-pose cascaded CNN for 2D hand pose estimation from single color images. *Transactions on Circuits and Systems for Video Technology (TCSVT)*, 29(11):3258–3268, 2019. 2

[102] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. DOPE: Distillation of part experts for whole-body 3D pose estimation in the wild. In *European Conference on Computer Vision (ECCV)*, pages 380–397, 2020. 3

[103] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10974, 2019. 3, 5, 6

[104] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020. 1, 3

[105] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 598–607, 2020. 1

[106] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018. 2

[107] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3D human mesh regression with dense correspondence. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7052–7061, 2020. 2

[108] Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5484–5493, 2017. 5

[109] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3D human shape and pose from dense body parts. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 8

[110] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocu-

lar RGB image. In *International Conference on Computer Vision (ICCV)*, pages 2354–2364, 2019. 1, 3

[111] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3D human pose regression. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3425–3435, 2019. 2

[112] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D human reconstruction from a single image. In *International Conference on Computer Vision (ICCV)*, pages 7738–7748, 2019. 2

[113] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 3

[114] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular Realtime Full Body Capture with Inter-part Correlations. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 7, 8

[115] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *International Conference on Computer Vision (ICCV)*, pages 4913–4921, 2017. 3

[116] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 5, 7