Neural Descent for Visual 3D Human Pose and Shape

Andrei Zanfir Eduard Gabriel Bazavan Mihai Zanfir
William T. Freeman Rahul Sukthankar Cristian Sminchisescu

Google Research

{andreiz, egbazavan, mihaiz, wfreeman, sukthankar, sminchisescu}@google.com

Abstract

We present deep neural network methodology to reconstruct the 3d pose and shape of people, including hand gestures and facial expression, given an input RGB image. We rely on a recently introduced, expressive full body statistical 3d human model, GHUM, trained end-to-end, and learn to reconstruct its pose and shape state in a self-supervised regime. Central to our methodology, is a learning to learn and optimize *approach, referred to as HUman Neural Descent (***HUND***), which avoids* both *second-order differentiation when* training the model parameters, and *expensive* state gradient descent *in order to accurately minimize a semantic differentiable rendering loss at test time. Instead, we rely on novel recurrent stages to update the pose and shape parameters such that not only losses are minimized effectively, but the process is meta-regularized in order to ensure end-progress.* **HUND***'s symmetry between training and testing makes it the first 3d human sensing architecture to natively support different operating regimes including self-supervised ones. In diverse tests, we show that* **HUND** *achieves very competitive results in datasets like H3.6M and 3DPW, as well as good quality 3d reconstructions for complex imagery collected in-the-wild.*

1. Introduction

Automatic 3d human sensing from images and video would be a key, transformative enabler in areas as diverse as clothing virtual apparel try-on, fitness, personal well-being, health or rehabilitation, AR and VR for improved communication or collaboration, self-driving systems with emphasis to urban scenarios, special effects, human-computer interaction or gaming, among others. Applications in shopping, telepresence or fitness would increase human engagement and stimulate collaboration, communication, and the economy, during a lock-down.

The rapid progress in 3D human sensing has recently relied on volumetric statistical human body models [24, 43] and supervised training. Most, if not all, state of the art architectures for predicting 2d, *e.g.*, body keypoints [5] or 3d, *e.g.*, body joints, kinematic pose and shape [30, 48, 16, 19, 36, 8, 17, 20, 2, 44, 18, 40, 15, 29, 45, 49, 33, 37, 27, 26, 14] rely, *ab initio*, at their learning core, on complete supervision. For 2d methods this primarily enters as keypoint or semantic segmentation annotations by humans, but for complex 3D articulated structures human annotation is both impractical and inaccurate. Hence for most methods, supervision comes in the form of synchronous 2d *and* 3d ground truth, mostly available in motion capture datasets like Human3.6M [13] and more recently also 3DPW [41].

Supervision-types aside, the other key ingredient of any successful system is the interplay between 3d initialization using neural networks and non-linear optimization (refinement) based on losses computed over image primitives like keypoints, silhouettes, or body part semantic segmentation maps. No existing feedforward system, particularly a *monocular* one, achieves *both* plausible 3d reconstruction *and* veridical image alignment[1] without non-linear optimization – a key component whose effectiveness for 3d pose estimation has been long since demonstrated [34, 35].

The challenge faced by applying non-linear optimization in high-dimensional problems like 3d human pose and shape estimation stems from its complexity. On one hand, first-order model state updates are relatively inefficient for very ill-conditioned problems like *monocular* 3d human pose estimation where Hessian condition numbers in the 10^{-3} are typical [34]. Consequently, many iterations are usually necessary for good results, even when BFGS approximations are used. On the other hand, nonlinear output state optimization is difficult to integrate as part of parameter learning, since correct back-propagation would require potentially complex, computationally expensive second-order

[1]To be understood in the classical model-based vision sense of best fitting the model predictions to implicitly or explicitly-associated image primitives (or landmarks), within modeling accuracy.

updates, for the associated layers. Such considerations have inspired some authors [19] to replace an otherwise desirable integrated learning process, with a dual system approach, where multiple non-linear optimization stages, supplying potentially improved 3d output state targets, are interleaved with classical supervised learning based on synchronized 2d and 3d data obtained by imputation. Such intuitive ideas have been shown to be effective practically, but remain expensive in training, and lack not just an explicit, integrated cost function, but also a consistent learning procedure to guarantee progress, in principle. Moreover, applying the system symmetrically, during testing, would still require potentially expensive non-linear optimization for precise image alignment.

In this paper, we take a different approach and replace the non-linear gradient refinement stage at the end of a classical 3d predictive architecture with neural descent, in a model called HUND (Human Neural Descent). In HUND, recurrent neural network stages refine the state output (in this case the 3d human pose and shape of a statistical GHUM model [43]) based on previous state estimates, loss values, and a context encoding of the input image, similarly in spirit to non-linear optimization. However, differently from models relying on gradient-based back-ends, HUND can be trained end-to-end using stochastic gradient descent, offers no asymmetry between training and testing, supports the possibility of potentially more complex, problem-dependent step updates compared to non-linear optimization, and is significantly faster. Moreover, by using such an architecture, symmetric in training and testing, with capability of refinement and self-consistency, we show, for the first time, that a 3d human pose and shape estimation system trained from monocular images can entirely bootstrap itself. The system would thus no longer necessarily require, the completely synchronous supervision, in the form of images and corresponding 3d ground truth configurations that has been previously unavoidable. Experiments in several datasets, ablation studies, and qualitative results in challenging imagery support and illustrate the main claims.

Related Work: There is considerable prior work in 3d human modeling [24, 43, 30, 48, 31, 16, 19, 7], as well as the associated learning and optimization techniques [34, 3]. Systems combining either random 3d initialization or prediction from neural networks with non-linear optimization using losses expressed in terms of alignment to keypoints and body semantic segmentation masks exist [3, 48, 19]. Black-box optimization has gained more interest in recent years [1, 6], usually deployed in the context of meta-learning [11]. Our work is inspired in part by that of [6, 11] in which the authors introduce recurrent mechanisms to solve optimization problems, albeit in a different domain and for other representations than the ones considered in this work. [28] uses a neural network to directly regress the pose and shape parameters of a 3d body model from predicted body semantic segmentation. The network is trained in a mixed supervision regime, with either full supervision for the body model parameters or a weak supervision based on a 2d reprojection loss. [42] propose to learn a series of linear regressors over SIFT [25] features that produce descent directions analogous to an optimization algorithm for face alignment. Training is fully supervised based on 2d landmarks. Similarly, [39] learn a recurrent network, that given an input image of a face, iteratively refines face landmark predictions. The network is trained fully supervised and operates only in the 2d domain. In [38], a cascade of linear regressors are learned to refine the 3d parameters of a 3d face model. Training is done over the entire dataset at a time (multiple persons with multiple associated face images) on synthetic data, in a simulated, mixed supervision regime.

2. Methodology

We describe the elements of the proposed methodology, including the statistical 3D human body model GHUM, as well as the associated learning and reconstruction architecture used. We also cover the fusion of multiple architectures in order to obtain accurate estimates for the full body including hand gestures and facial expressions.

2.1. Statistical 3D Human Body Model GHUM

We use a recently introduced statistical 3d human body model called GHUM [43], to represent the pose and the shape of the human body. The model has been trained end-to-end, in a deep learning framework, using a large corpus of over 60,000 diverse human shapes, and $540,000$ human motions, consisting of 390,000 samples from CMU and 150,000 samples from Human3.6M (subjects S1, S5, S6, S7, S8). The model has generative body shape and facial expressions $\boldsymbol{\beta} = (\boldsymbol{\beta_b}, \boldsymbol{\beta_f})$ represented using deep variational auto-encoders and generative pose $\boldsymbol{\theta} = (\boldsymbol{\theta_b}, \boldsymbol{\theta_{lh}}, \boldsymbol{\theta_{rh}})$ for the body, left and right hands respectively represented using normalizing flows [46]. We assume a separable prior on the model pose and shape state $p(\boldsymbol{\theta}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}) + p(\boldsymbol{\beta})$ where Gaussian components with $\mathbf{0}$ mean and unit $\mathbf{I}$ covariance, as typical in variational encoder and normalizing flow models. Given a monocular RGB image as input, our objective is to infer the pose $\boldsymbol{\theta} \in \mathbb{R}^{N_p \times 1}$ and shape $\boldsymbol{\beta} \in \mathbb{R}^{N_s \times 1}$ state variables, where N_p is the number of posing variables and N_s is the length of the shape code, respectively. A posed mesh $\mathbf{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ has N_v associated 3d vertices $\mathbf{V} = \{\mathbf{v}_i, i = 1 \dots N_v\}$ with fixed topology given by the GHUM template. Because the rigid transformation of the model in camera space – represented by a 6d rotation [50] $\mathbf{r} \in \mathbb{R}^{6 \times 1}$ and a translation vector $\mathbf{t} \in \mathbb{R}^{3 \times 1}$ – are important and require special handling, we will write them explicitly. The posed mesh thus writes $\mathbf{M}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{r}, \mathbf{t})$.

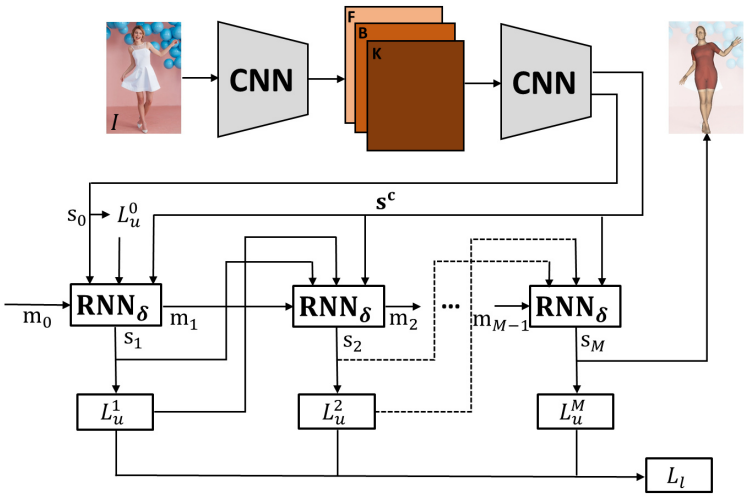

Figure 1. Overview of our Human Neural Descent (**HUND**) architecture for learning to estimate the state $\mathbf{s}$ of a generative human model GHUM (including shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$, as well as person's global rotation $\mathbf{r}$ and translation $\mathbf{t}$) from monocular images. Given an input image, a first CNN extracts semantic feature maps for body keypoints ($\mathbf{K}$) and part segmentation ($\mathbf{B}$), as well as other features ($\mathbf{F}$). These, in turn feed, into a second stage CNN that learns to compute a global context code $\mathbf{s}^c$ as well as an initial estimate of the model state $\mathbf{s}_0$. These estimates (and at later stages similar ones obtained recursively), together with the value of a semantic alignment loss L_u, expressed in terms of keypoint correspondences and differentiable rendering measures between model predictions and associated image structures, are fed into multiple refining RNN layers, with shared parameters $\boldsymbol{\delta}$, and internal memory (hidden state) $\mathbf{m}$. The alignment losses (which can be unsupervised, weakly-supervised or self-supervised, depending on available data) at multiple recurrent stages M are aggregated into a learning loss L_l, optimized as part of the learning-to-learn process. The parameters are obtained using stochastic gradient descent, as typical in deep learning. The model produces refined state estimates $\mathbf{s}$ with precise image alignment, but does not require additional gradient calculations for the recurrent stages neither in training (*e.g.*, second-order parameter updates), nor during testing (first-order state updates). It is also extremely efficient computationally compared to models relying on nonlinear state optimization at test time.

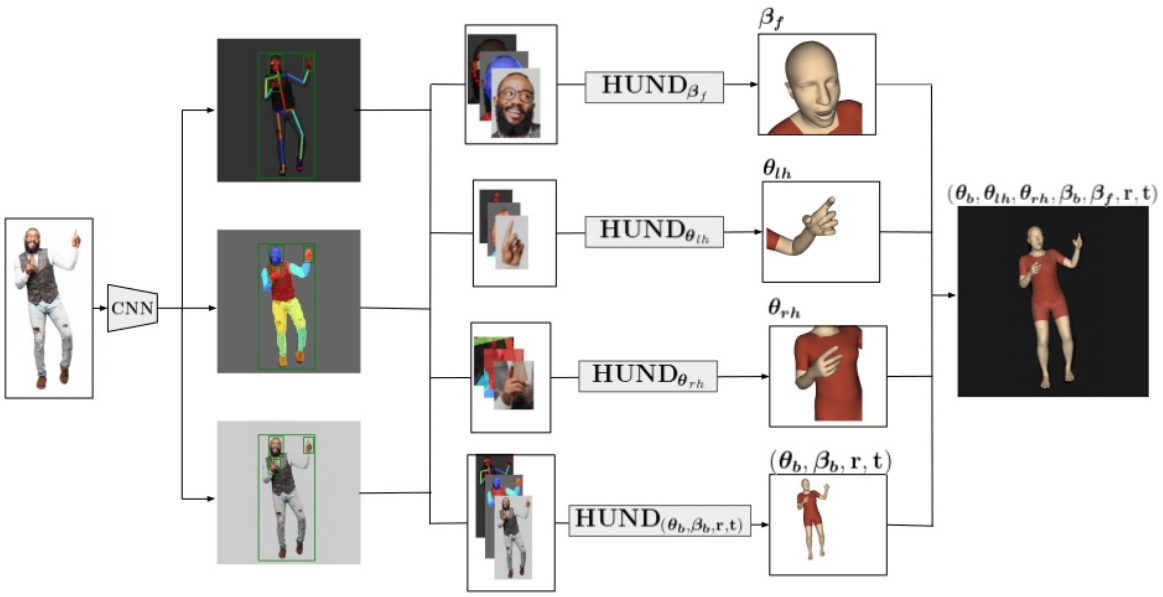

Figure 2. Our complete full body 3d sensing HUND network combines a feed-forward architecture to detect landmarks and semantically segment body parts with an attention mechanism that further processes the face, hands and the rest of the body as separate HUND predictive networks, with results fused in order to obtain the final, full body estimate. See fig.1 for the architecture of an individual HUND network.

Camera model. We assume a pinhole camera with intrinsics $C = [f_x, f_y, c_x, c_y]^\top$ and associated perspective projection operator $\mathbf{x}_{2d} = \Pi(\mathbf{x}_{3d}, C)$, where $\mathbf{x}_{3d}$ is any 3d point. During training and testing, intrinsics *for the full input image* are approximated, $f_x = \max(H, W), f_y = \max(H, W), c_x = W/2, c_y = H/2$, where H, W are the input dimensions. Our method works with cropped bounding-boxes of humans, re-scaled to a fixed size of 480×480, therefore we need to warp the image intrinsics C into the corresponding crop intrinsics C_c

$$[C_c^\top 1]^\top = K[C^\top 1]^\top, \tag{1}$$

where $K \in \mathbb{R}^{5\times5}$ is the scale and translation matrix, adapting the image intrinsics C. By using cropped intrinsics, we effectively solve for the state of the 3d model (including global scene translation) in the camera space of the input image. For multiple detections in the same image, the resulting 3d meshes are estimated relative to a common world coordinate system, into the *same 3d scene*. At test time, when switching C_c with C, the 3d model projection will also align with the corresponding person layout in the initial image.

2.2. Learning Architecture

The network takes as input a cropped human detection and resizes it to 480×480. A multi-stage sub-network produces features $\mathbf{F} \in \mathbb{R}^{60\times60\times256}$, keypoint detection heatmaps $\mathbf{K} \in \mathbb{R}^{60\times60\times137}$ and body-part segmentation maps $\mathbf{B} \in \mathbb{R}^{60\times60\times15}$. These are embedded into a low-dimensional space, producing a code vector $\mathbf{s}^c$ – the superscript c stands for context, *i.e.* the optimization's objective function context. We also append the cropped camera intrinsics C_c to this context vector. At training time, a estimate $\mathbf{s}_0$ of the initial GHUM state $\mathbf{s} = [\boldsymbol{\theta}^\top, \boldsymbol{\beta}^\top, \mathbf{r}^\top, \mathbf{t}^\top]^\top$ is also produced. To simulate model refinement[2], we employ a Recurrent Neural Network module $\mathbf{RNN}_{\boldsymbol{\delta}}(\mathbf{s}^c, \mathbf{s}_i, \mathbf{m}_i)$, where $\mathbf{m}_i$ is the memory (hidden state) at refinement stage i, and unroll the updates into M stages (see fig.1)

$$\begin{bmatrix}\mathbf{s}_i \\ \mathbf{m}_i\end{bmatrix} = \mathbf{RNN}_{\boldsymbol{\delta}}(\mathbf{s}_{i-1}, \mathbf{m}_{i-1}, L_u^{i-1}, \mathbf{s}^c). \tag{2}$$

The loss at each stage i is computed based on the labeling available at training time in the form of either 2d or 3d annotations. When both are missing, we are training with *self-supervision*. The self-supervised loss at each unit processing stage i can be expressed as

$$L_u^i(\mathbf{s}, \mathbf{K}, \mathbf{B}) = \lambda_k L_k(\mathbf{s}_i, \mathbf{K}) + \lambda_b L_b(\mathbf{s}_i, \mathbf{B}) + l(\boldsymbol{\theta}_i, \boldsymbol{\beta}_i), \tag{3}$$

where $l = -\log(p)$, L_k is a 2d keypoint alignment loss, L_b is a 2d semantic body part alignment (defined in terms of differentiable rendering), and M is the total number of *training* LSTM stages, while λ_k and λ_b are cross-validated scalar values which balance the loss terms.

The **keypoint alignment loss**, L_k, measures the reprojection error of the GHUM's model 3d joints w.r.t. the predicted 2d keypoints. The loss is defined as the 2d mean-per-joint position error (MPJPE)

$$L_k(\mathbf{s}_t, \mathbf{K}) = \frac{1}{N_j}\sum_i^{N_j} \|\mathbf{j}_i(\mathbf{K}) - \Pi(\mathbf{J}_i(\mathbf{s}_t), C_c)\|_2. \tag{4}$$

with N_j keypoints, $\mathbf{j}_i(\mathbf{K})$ is the 2d location of the i-th 2d keypoint extracted from the the $\mathbf{K}$ heatmap, and $\mathbf{J}_i(\mathbf{s}_t)$ is the i-th 3d keypoint computed by posing the GHUM model at $\mathbf{s}_t$.

The **body-part alignment loss**, L_b, uses the current prediction $\mathbf{s}_t$ to create a body-part semantic segmentation image $I(\mathbf{M}(\mathbf{s}_t), C_c) \in \mathbb{R}^{H\times W\times15}$. Then we follow a soft differentiable rasterization process[23] to fuse probabilistic contributions of all predicted mesh triangles of the model, at its current state, with respect to the rendered pixels. In this way, gradients can flow to the occluded and far-range vertices. To be able to aggregate occlusion states and semantic information, we append to each mesh vertex its semantic label, as a one-hot vector $\{0, 1\}^{15\times1}$, and a constant alpha value of 1. The target body part semantic probability maps $\mathbf{B}$ are also appended with a visibility value, equal to the foreground probability $\in [0, 1]^{H\times W\times1}$. The loss is the mean-per-pixel absolute value of the difference between the estimated and predicted semantic segmentation maps

$$L_b(\mathbf{s}_t, \mathbf{B}) = \frac{1}{HW}\sum_i^{HW} \|\mathbf{B}_i - I(\mathbf{M}(\mathbf{s}_t), C_c)_i\|_1. \tag{5}$$

For **body shape and pose**, we include two regularizers, proportional to the negative log-likelihood of their associated Gaussian distributions

$$l(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2, \;\; l(\boldsymbol{\beta}) = -\log p(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2. \tag{6}$$

When *3d supervision* is available, we use the following unit training loss L_f^i, as well as, potentially, the other ones previously introduced in (3) for the self-supervised regime

$$L_f^i(\mathbf{s}) = \lambda_m L_m(\mathbf{M}(\mathbf{s}_i), \widetilde{\mathbf{M}}) + \lambda_{3d} L_{3d}(\mathbf{J}(\mathbf{s}_i), \widetilde{\mathbf{J}}),$$

[2]HMR[16] uses several recursive output layers on top of a CNN prediction. However HMR does not use a formal RNN to recursively refine outputs based on a memory structure encompassing the previous estimates, the image reprojection (keypoint and semantic) error and the image feature code, as we do, which is the equivalent of a complete non-linear optimization context. Nor do we use a discriminator for pose as HMR, but instead rely on the kinematic normalizing flow prior of GHUM. Hence our approach is methodologically very different. See §3 for quantitative evaluation.

where L_m represents the 3d vertex error between the ground-truth mesh $\widetilde{\mathbf{M}}$ and a predicted one, $\mathbf{M}(\mathbf{s}_i)$– obtained by posing the GHUM model using the predicted state $\mathbf{s}_i$; L_{3d} is the 3d MPJPE between the 3d joints recovered from the predicted GHUM parameters, $\mathbf{J}(\mathbf{s}_i)$, and the ground-truth 3d joints, $\widetilde{\mathbf{J}}$; λ_m and λ_{3d} are scalar values that balance the two terms.

For learning, we consider different losses L_l, including 'sum', 'last', 'min' or 'max', as follows

$$
\begin{aligned}
L_u^{\Sigma}(\mathbf{s}, \mathbf{K}, \mathbf{B}) &= \sum_{i=1}^{M} L_u^i(\mathbf{s}_i, \mathbf{K}, \mathbf{B}) \\
L_u^{\rightarrow}(\mathbf{s}, \mathbf{K}, \mathbf{B}) &= L_u^M(\mathbf{s}_M, \mathbf{K}, \mathbf{B}) \\
L_u^{\min}(\mathbf{s}, \mathbf{K}, \mathbf{B}) &= \min_{i=1}^{M} L_u^i(\mathbf{s}_i, \mathbf{K}, \mathbf{B}) \\
L_u^{\max}(\mathbf{s}, \mathbf{K}, \mathbf{B}) &= \max_{i=1}^{M} L_u^i(\mathbf{s}_i, \mathbf{K}, \mathbf{B})
\end{aligned} \tag{7}
$$

We also consider an *observable improvement* (OI) loss for L_l [11]

$$
L_u^{oi} = \sum_{i=1}^{M} \min\{L_u^i - \min_{j<i} L_u^j, 0\}. \tag{8}
$$

Multiple HUND networks for Body Pose, Shape, and Facial Expressions. Capturing the main body pose but also hand gestures and facial expressions using a single network is challenging due to the very different scales of each region statistics. To improve robustness and flexibility we rely on 4 part networks, one specialized for facial expressions, two for the hands, and one for the rest of the body. Based on an initial person keypoint detection and semantic segmentation, we drive attention to face and hand regions as identified by landmarks and semantic maps, in order to process those features in more detail. This results in multiple HUND networks being trained, with estimates for the full body shape and pose fused in a subsequent step from parts (fig. 2).

3. Experiments

View of Experimental Protocols. There is large variety of models and methods now available for 3d human sensing research, including body models like SMPL [24] and GHUM [43], or reconstruction methods like DMHS [30], HMR [16], SPIN [19] *etc*., set aside methods that combine random initialization or neural network prediction and non-linear refinement[3, 47]. To make things even more complex, some models are pre-trained on different 2d or 3d datasets and refined on others. A considerable part of this development has a historical trace, with models built on top of each-other and inheriting their structure and training sets, as available at different moments in time. Set that aside, multiple protocols are used for testing. For Human3.6M [13] only, there are at least 4: the ones originally proposed by the dataset creators, on the withheld test set of Human3.6M (or the representative subset Human80K [12]) as well as others, created by various authors, known as protocol 1 and 2 by re-partitioning the original training and validation sets for which ground truth is available. Out of these 2, only protocol 1 is sufficiently solid in the sense of providing a reasonably large and diverse test set for stable statistics (*e.g*., 110,000 images from different views in P1 vs. 13,700 in P2, from the same camera, at the same training set size of 312,000 configurations for both). Hence we use P1 for ablations and the official Human3.6M test set for more relevant comparisons. For some of the competing methods, e.g. SPIN[19], HMR[16] we ran the code from the released github repositories ourselves on the Human3.6M test set since numbers were not reported in the original publications. Results are presented in table 2. We will also use 3DPW [41] for similar reasons, or rather, in the absence of other options in the wild (30,150 training and 33,000 testing configurations). Testing all other model combinations would be both impractical and irrelevant, especially for new models like GHUM where most prior combinations are unavailable and impossible to replicate. As a matter of principle, 3D reconstruction models can be evaluated based on the amount of supervision received, be it 2d (for training landmark detectors **#2d det** or, additionally, for direct 3d learning **#2d**), **# 3d**, or synchronized **#2d-3d** annotations, the number of images used for self-supervision #I, as well as perhaps number of parameters and run-time. In addition, ablations for each model, *e.g*., **HUND**, would offer insights into different components and their relevance. We argue in support of this being one scientifically sound way of promoting diversity in the creation of new models and methods, rather than closing towards premature methodological convergence, weakly supported by unsustainable, ad-hoc, experimental combinatorics.

For our **self-supervised (SS)** experiments, we employ two datasets containing images in-the-wild, COCO2017 [22] (30,000 images) and OpenImages [21] (24,000), with no annotations in training and testing. We refer to **weakly-supervised (WS)** experiments as those where ground truth annotations are available, *e.g*. human body keypoints. We do not rely on these but some other techniques including HMR and SPIN do, hence we make this distinction in order to correctly reflect their supervision level.

For **fully supervised (FS)** experiments, we employ H3.6M and 3DPW. Because we work with the newly released GHUM model, we retarget the mocap raw marker data from H3.6M to obtain accurate 3d mesh supervision for our model [43]. Because the ground-truth of 3DPW is provided as SMPL 3d meshes, we fit the GHUM model by using an objective function minimizing vertex-to-vertex distances between the two corresponding meshes.

Architecture Implementation. To predict single-person

Method	MPJPE-PA	MPJPE	MPJPE Trans	#2d det	#2d	#3d	#2d-3d	#I
HMR (FS+WS) [16]	58.1	88.0	NR	129k	111k	720k	300k	0
SPIN (FS+WS) [19]	NR	NR	NR	129k	111k	720k/390k	300k	0
HUND (FS+SS)	**52.6**	**69.45**	152.6	80k	0	540k	150k	54k
HMR (WS) [16]	67.45	106.84	NR	129k	111k	720k	0	0
HUND (SS)	**66.0**	**91.8**	159.3	80k	0	540k	0	54k

Table 1. Performance of different pose and shape estimation methods on the H3.6M dataset, with training/testing based on the representative protocol P1 (for self-supervised variants this only indicates the images used in testing). MPJPE-PA and MPJPE are expressed in mm. We also report the global translation of the body as this is supported by our fully perspective camera model (N.B. this is not supported by other methods which use an orthographic perspective model). We also compare different annotations used in the construction of different models, with a split into 2d (further differentiated into **#2d det** for training the joint landmarks and **#2d** for training the 3d learning algorithm), 3d and synchronized 2d-3d. The last column gives the number of images for self-supervised variants, *e.g.*, HUND(SS), which do not use either 2d image keypoints or synchronized images and 3d mocap during training.

Method	MPJPE
HMR (FS+WS) [16]	89
SPIN (FS+WS) [19]	68
HUND (FS+SS)	**66**

Table 2. Results of different methods on the H3.6M official held-out test set. We achieve better results on a large test set of 900k images.

Method	MPJPE-PA (mm)	MPJPE (mm)
HMR (FS+WS) [16]	81.3	130.0
SPIN (FS+WS)[19]	59.2	96.9
ExPose (FS+WS)[7]	60.7	93.4
HUND (SS)	63.5	90.4
HUND (FS+SS)	**57.5**	**81.4**

Table 3. Results on the 3DPW test set for different methods. Notice that a self-supervised version of HUND produces lower errors compared to the best supervised HMR implementation that includes not just synchronized $2d-3d$ training sets but also images with 2d annotation ground truth. A HUND model that includes asynchronous 2d-3d supervision, in addition to just unlabeled images, achieves the lowest error, and uses less training data than any other competitive method – see also table 1.

keypoints and body part segmentation, we train a multi-task network with ResNet50 [9] backbone (the first CNN in our pipeline, see fig.1 and fig.2)[30]. We have 137 2d keypoints as in [4] and 15 body part labels, as in [32]. This network has 34M trainable parameters. The self-supervised training protocol for **HUND** assumes only images are available and we predict 2d body keypoints and body part labels during training, in addition to body shape and pose regularizers. For the embedding model (the second CNN in the pipeline) predicting $\mathbf{s}^c$, we use a series of 6 convolutional layers with pooling, followed by a fully connected layer. We use $M = 5$ LSTM [10] stages as RNNs for **HUND** and we set the number of units to 256, which translates to 525k parameters. In total there are 950k trainable parameters for the 3D reconstruction model. We train with a batch size of 32 and a learning rate of 10^{-4} for 50 epochs. For experiments where we train HUND using **FS+SS**, we use a mixed schedule, alternating between self-supervised and fully-supervised batches. Training takes about 72 hours on a single Nvidia Tesla P100 GPU. The runtime of our prediction network for a single image is 0.035s and 0.02s for **HUND**, on an Nvidia RTX 2080 GPU.

Evaluation and discussion. Multiple experiments are run for different regimes. Quantitative results are presented in tables 1, 2 and 3 for Human3.6M and 3DPW respectively. A detailed analysis of optimization behavior for one image is given in fig. 3 as well as, in aggregate, in fig. 4. Visual reconstructions at different HUND optimization stages, for several images, are given in fig. 6.

Loss	MPJPE-PA (mm)	MPJPE (mm)
$L_f^{\rightarrow}$	**58.50**	**80.16**
L_f^{Σ}	59.91	83.26
$L_f^{\min}$	78.61	122.60
L_f^{oi}	79.35	123.90
$L_f^{\max}$	83.80	128.0

Table 4. Impact assessment of different meta-losses used in HUND (FS), trained on the Human3.6M dataset, following protocol 1. The last and sum losses perform similarly well, with others following at a distance.

We also study the impact of different meta-learning losses, as given in (7) and (8), on the quality of results of HUND. We use a HUND (FS) model trained and evaluated on Human3.6M (protocol 1). From table 4 we observe that the last ($L_f^{\rightarrow}$) and sum (L_f^{Σ}) losses perform best, whereas others produce considerably less competitive results, by some margin, for this problem. Finally, we show qualitative visual 3d reconstruction results, from several viewpoints, for a variety of difficult poses and backgrounds in fig. 5. *Please see our Sup. Mat. for videos!*

Ethical Considerations. Our methodology aims to decrease bias by introducing flexible forms of self-supervision which would allow, in principle, for system bootstrapping and adaptation to new domains and fair, diverse subject distributions, for which labeled data may be difficult or impossible to collect upfront. Applications like visual surveillance and person identification would not be effectively supported currently, given that model's output does not provide sufficient

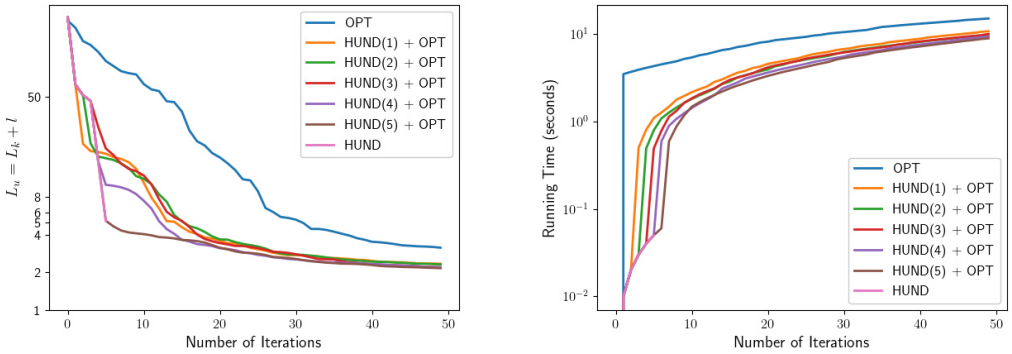

Figure 3. Behavior of different optimization methods including standard non-linear gradient-based BFGS, HUND(5), as well as variants of HUND$(i), i \leq 5$, initializing BFGS, in order to assess progress and the quality of solutions obtained along the way (left). Corresponding cumulative run-times are shown on the right. Observe that HUND produces a good quality solution orders of magnitude faster than gradient descent (note log-scales on both plots). End refinement using gradient descent improves results, although we do not recommend a hybrid approach— here we only show different hybrids for insight. This shows one optimization trace for a model initialized in A-pose and estimated given one image from Human3.6M, but such behavior is typical of aggregates, see *e.g.*, fig. 4. See also fig. 6 for visual illustrations of different configurations sampled by HUND during optimization.

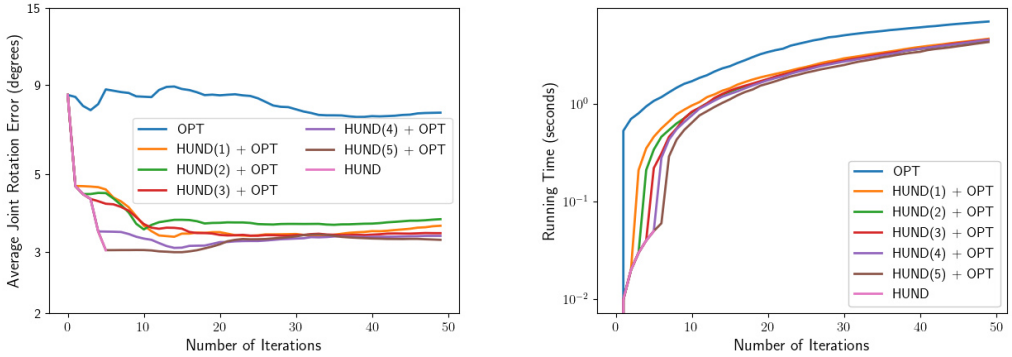

Figure 4. Optimization statistics for different methods, aggregated over 100 different poses (estimation runs) from Human3.6M. We initialize in an A-pose and perform monocular 3d pose and shape reconstruction for GHUM under a HUND (FS+SS) model, as well as non-linear optimization baselines. On the left we show per-joint angle averages w.r.t. ground truth. On the right we show running times in aggregate for different types of optimization. One can see that BFGS descent under a keypoint+prior loss tends to be prone to inferior local optima compared to different HUND hybrids, which on average find significantly better solutions. The plot needs to be interpreted in proper context, as aggregates meant to show distance and run-time statistics per iteration. Hence, they may not be entirely representative of any single run, but for a singleton see *e.g.*, fig. 3.

detail for these purposes. This is equally true of the creation of potentially adversely-impacting deepfakes, as we do not include an appearance model or a joint audio-visual model.

4. Conclusions

We have presented a neural model, **HUND**, to reconstruct the 3d pose and shape of people, including hand gestures and facial expressions, from image data. In doing so, we rely on an expressive full body statistical 3d human model, GHUM, to capture typical human shape and motion regularities. Even so, accurate reconstruction and continuous learning are challenging because large-scale diverse 3d supervision is difficult to acquire for people, and because the most efficient inference is typically based on non-linear image fitting. This is however difficult to correctly 'supra'-differentiate, to second order, in training and expensive in testing. To address such challenges, we rely on self-supervision based on differentiable rendering within *learning-to-learn* approaches based on recurrent networks, which avoid expensive gradient descent in testing, yet provide a surrogate for robust loss minimization. **HUND** is tested and achieves very competitive results for datasets like H3.6M and 3DPW, as well as for complex poses, collected in challenging outdoor conditions. HUND's learning-to-learn and optimize capabilities, and symmetry between training and testing, can make it the first architecture to demonstrate the possibility of bootstraping a plausible 3d human reconstruction model without initial, synchronous (2d, 3d) supervision.

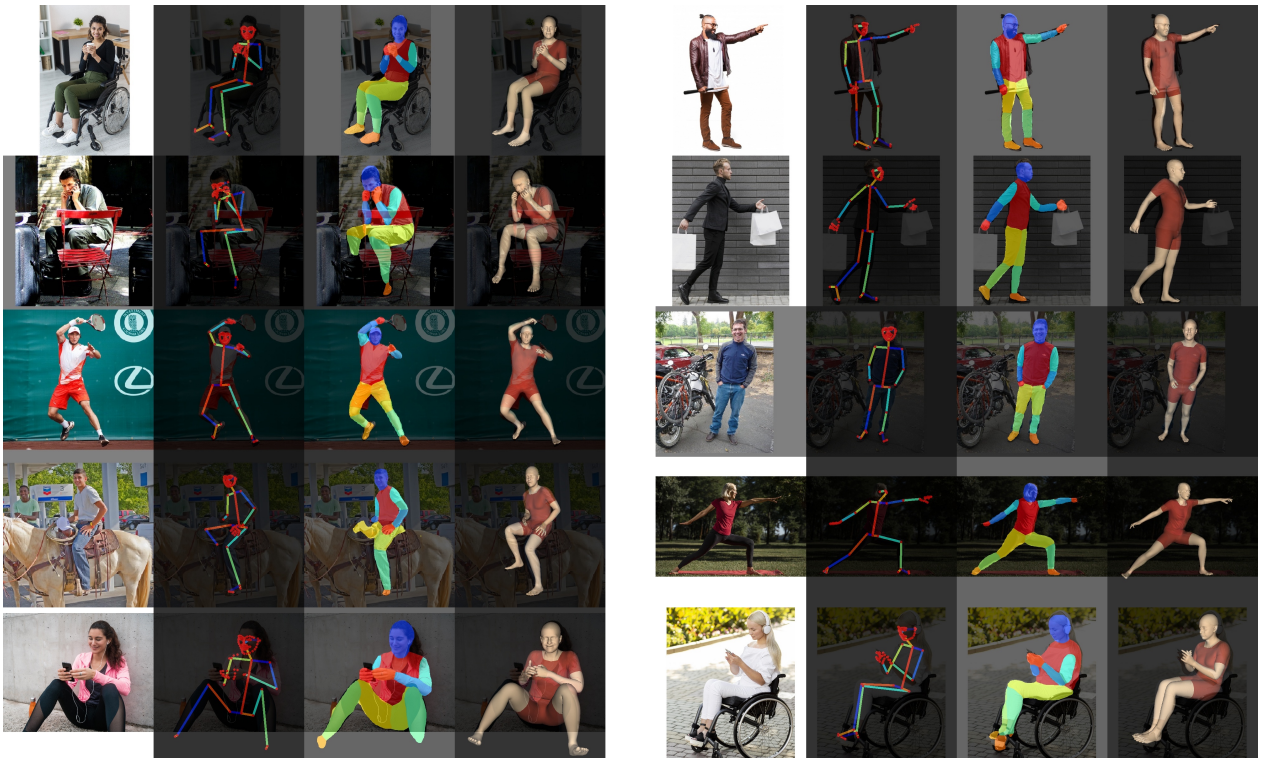

Figure 5. Visual 3d reconstruction results obtained by **HUND**. Given initial 2d predictions for body, face and hand keypoints, and initial predictions for semantic body part labelling, the neural descent network predicts the 3d GHUM pose and shape parameters. Best seen in color. For other examples and videos see our Sup. Mat.

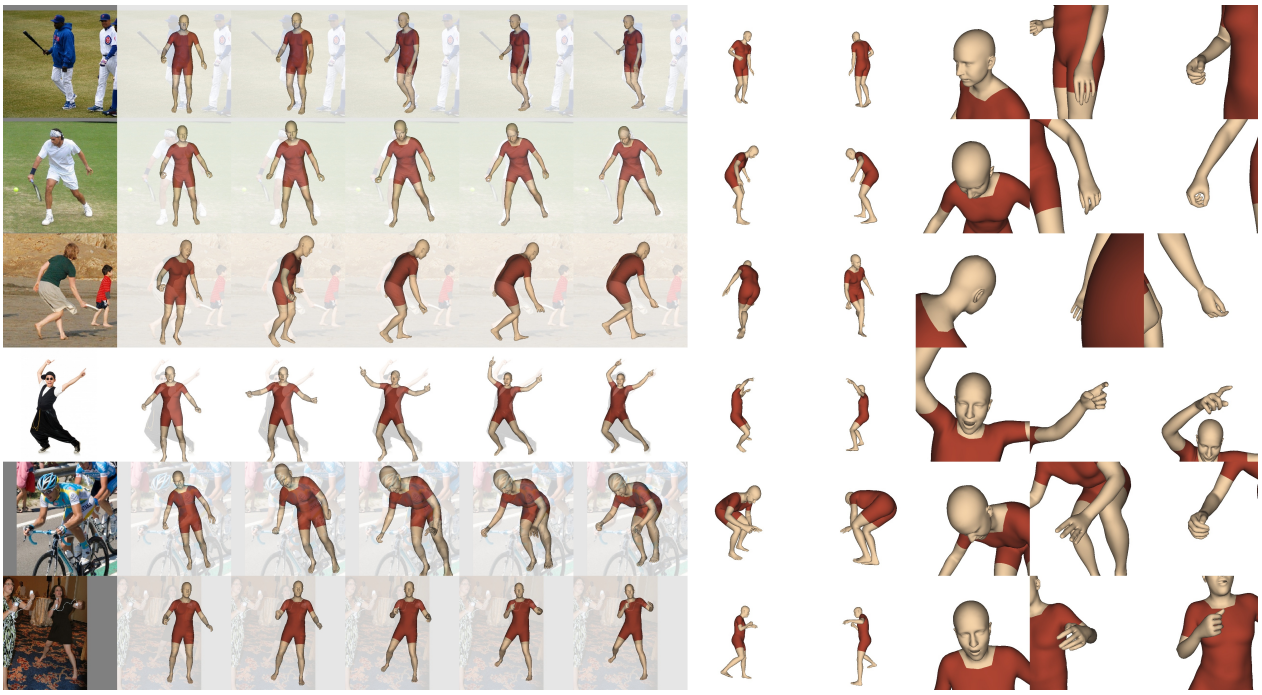

Figure 6. Visual 3d pose and shape configurations of GHUM sampled by **HUND** during optimization. First column shows the input image, columns 2-6 illustrate GHUM estimates at each HUND stage. Columns 7 and 8 show visualizations of the GHUM state from different viewpoints, after HUND terminates. Columns 9, 10 and 11 show close up views for the reconstructed face expressions, left and right hands.

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *NeurIPS*, pages 3981–3989, 2016. 2

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild, 2019. 1

[3] 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 *ECCV*, 2016. 2, 5

[4] Z Cao, G Martinez Hidalgo, T Simon, SE Wei, and YA Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *PAMI*, 2019. 6

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1

[6] Yutian Chen, Matthew W Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Timothy P Lillicrap, Matt Botvinick, and Nando De Freitas. Learning to learn without gradient descent by gradient descent. In *ICML*, pages 748–756, 2017. 2

[7] 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)*, 2020. 2, 6

[8] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In *Advances in Neural Information Processing Systems*, pages 12929–12941, 2019. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016. 6

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[11] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey, 2020. 2, 5

[12] Catalin Ionescu, João Carreira, and Cristian Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *CVPR*, 2014. 5

[13] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014. 1, 5

[14] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision (ICCV)*, 2019. 1

[15] Aaron S Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *ECCV*, 2018. 1

[16] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 4, 5, 6

[17] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. *CVPR*, 2020. 1

[19] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 1, 2, 5, 6

[20] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 1

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018. 5

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[23] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *arXiv preprint arXiv:1904.01786*, 2019. 4

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH*, 2015. 1, 2, 5

[25] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[26] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 1

[27] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 2017. 1

[28] 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 *3DV*, 2018. 2

[29] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017. 1

[30] A.I. Popa, M. Zanfir, and C. Sminchisescu. Deep Multitask Architecture for Integrated 2D and 3D Human Sensing. In *CVPR*, 2017. 1, 2, 5, 6

[31] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *ECCV*, September 2018. 2

[32] Iasonas Kokkinos Riza Alp Guler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. *arXiv*, 2018. 6

[33] Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *NeurIPS*, 2016. 1

[34] C. Sminchisescu and B. Triggs. Estimating Articulated Human Motion with Covariance Scaled Sampling. *IJRR*, 22(6):371–393, 2003. 1, 2

[35] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *CVPR*, 2003. 1

[36] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5349–5358, 2019. 1

[37] Bugra Tekin, Pablo Marquez Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *ICCV*, 2017. 1

[38] Wan Tian, Feng Liu, and Qijun Zhao. Regressing 3d face shapes from arbitrary image sets with disentanglement in shape space. In *2019 International Conference on Biometrics (ICB)*, pages 1–7. IEEE, 2019. 2

[39] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, pages 4177–4187, 2016. 2

[40] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 1

[41] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018. 1, 5

[42] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013. 2

[43] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. *CVPR*, 2020. 1, 2, 5

[44] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7760–7770, 2019. 1

[45] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 1

[46] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. *ECCV*, 2020. 2

[47] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes – The Importance of Multiple Scene Constraints. In *CVPR*, 2018. 5

[48] 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 *CVPR*, 2018. 1, 2

[49] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 1

[50] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *arXiv preprint arXiv:1812.07035*, 2018. 2