Weakly Supervised 3D Human Pose and Shape Reconstruction with Normalizing Flows

Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu
William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu

Google Research
{andreiz,egbazavan,hongyixu,wfreeman,sukthankar,sminchisescu}@google.com

Abstract. Monocular 3D human pose and shape estimation is challenging due to the many degrees of freedom of the human body and the difficulty to acquire training data for large-scale supervised learning in complex visual scenes. In this paper we present practical semi-supervised and self-supervised models that support training and good generalization in real-world images and video. Our formulation is based on kinematic latent normalizing flow representations and dynamics, as well as differentiable, semantic body part alignment loss functions that support self-supervised learning. In extensive experiments using 3D motion capture datasets like CMU, Human3.6M, 3DPW, or AMASS, as well as image repositories like COCO, we show that the proposed methods outperform the state of the art, supporting the practical construction of an accurate family of models based on large-scale training with diverse and incompletely labeled image and video data.

Keywords: 3D human sensing, normalizing flows, semantic alignment.

1 Introduction

Recovering 3D human pose and shape from monocular RGB images is important for motion and behavioral analysis, robotics, self-driving cars, computer graphics, and the gaming industry. Considerable progress has been made recently in increasing the size of datasets, in the level of detail of human body modeling, and the use of deep learning. A difficulty is the somewhat limited diversity of supervision available in the 3D domain. Many datasets offer 2D human body joint annotations or semantic body part segmentation masks for images collected in the wild, but lack 3D annotations. Motion capture datasets in turn offer large and diverse 3D annotations but their image backgrounds, clothing or body shape variation is not as high. Multi-task models, or models able to learn using limited forms of supervision, represent a potential solution to the current 3D supervision limitations. However, the number of human body shapes and poses observed in images collected in the wild is large, so strong pose, shape priors and expressive loss functions appear necessary in order to make learning feasible. In this paper we address some of these challenges by designing a family of normalizing flow based kinematic priors, together with semantic alignment

losses that make large scale weakly and self-supervised learning more accurate and efficient. *The introduction and integration of these components, new in the framework of human sensing, with strong results, is one of the main contributions of this work.* An evaluation (with ablation studies) on large scale datasets like Human3.6M, COCO, 3DPW, indicates good weakly supervised performance for 3D reconstruction. Our proposed priors and loss functions are amenable to both integration into deep learning losses and to direct non-linear state optimization (refinement) of a model given a random seed or initialization from a learnt predictor.

Mindset. Our use of different data sources is practically minded, as we aim towards large scale operation in the wild. Hence we rely on all types of supervision and data sources available. We often start with models trained in the lab, e.g. using Human3.6M and those are *supervised*. We also use 3D motion capture repositories like CMU in order to construct kinematic (output) priors and that component alone would make our approach *semi-supervised*. Finally, we make use of large scale predict-and-reproject losses for unlabeled datasets like MS COCO, which makes our approach, at least to an extent *self-supervised*. Whatever model curriculum used, we aim, long-term, to converge on self-supervised operation. We work with a semi-supervised output prior and model ignition is based on supervision in the lab. By convention, we call this regime *weakly-supervised*.

Related Work. There is considerable work in 3D human pose estimation based on 2D keypoints, semantic segmentation of body parts, and 3D joint positions [36, 38, 27, 33, 26, 9]. More recently, there has been significant interest in 3D human pose and shape estimation [37, 6, 13, 20, 1, 43], with some in the form of a reduced parametric model [31] decoded by 2D predictions, volumetric variants [40] or direct vertex prediction combined with 3D model fitting [10, 45]. Learning under weak supervision represents the next frontier, considered in this work as well. [44] learns a discriminator in order to transfer knowledge gained on a 3D dataset to a 2D one. [47] train a shared representation for both 2D and 3D pose estimation, with a regularizer operating on body segments in order to preserve statistics. [12] use a discriminator as prior, with adversarial training, and mixes 3D supervision and image labels. [28] uses segmentations as an intermediate layer, defines a loss on 2D and 3D joints, and rely on rotation matrices instead of angle-axiss representation. [32] uses a differentiable renderer (OpenDR) to compute a silhouette loss with a limited basin of attraction. This is only used for finetuning the network, but the authors report not having observed significant gains. [39] rely on a segmentation loss defined on silhouettes, not on the body parts, and rely on multiple views and temporal constraints for learning.

A variety of methods rely on priors for 3D optimization starting from an initial estimate provided by a neural network and/or by relying on image features like keypoints or silhouettes. [2] fit a Gaussian mixture model to motion capture data from CMU [11] and use it during optimization. We will evaluate this prior in our work. SPIN[19] alternates rounds of training with estimation of new targets using optimization (we will compare in §3).

Multiple differentiable rendering models [15, 23, 35] have been proposed recently, in more general settings. Such models are elegant and offer the promise of optimizing photo-realistic losses in the long run. The challenge is in defining an end-to-end model that embeds the difficult assignment problem between the model predictions (rasterized or not) and the image features in ways that are both differentiable and amenable to larger basins of attraction. Our semantic alignment loss is not technically a rendering model, but is differentiable and offers large basins of attraction, supported by explicit, long-range semantic body part correspondences. Gradients can be propagated for points that are not rendered (i.e. points that fail the z-test) and the operation is parallelizable and easy to implement.

2 Methodology

3D Pose and Shape Representations. We use a statistical body model [22, 42] to represent the pose and the shape of the human body. Given a monocular RGB image, our objective is to infer the pose state variables $\boldsymbol{\theta} \in \mathbb{R}^{N_j \times 3}$ and shape $\boldsymbol{\beta} \in \mathbb{R}^{N_s}$. A posed mesh $\mathbf{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ has N_v associated 3D vertices $\mathbf{V} = \{\mathbf{v}_i, i = 1 \dots N_v\}$. By dropping dependency on parameters we sometimes denote $\mathbf{M}(\mathbf{V}, k)$ the subset of vertices associated with body part index k (e.g. torso or head).

For prediction and optimization tasks we experiment with several kinematic representations. The angle-axis gives good results in connection with deep learning architectures [12, 45]. The representation consists of a set of N_j angle-axis variables $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_{N_j}\}, \boldsymbol{\theta}_i \in \mathbb{R}^3$, where the norm of $\boldsymbol{\theta}_i$ is the rotation angle in radians and $\frac{\boldsymbol{\theta}_i}{\|\boldsymbol{\theta}_i\|}$ is the unit length 3D axis of rotation.

We also explore a new 6D over-parameterization of rotations [48], given by the first two columns of the rotation matrix. We test this parameterization in the context of optimization, by building a prior and minimizing a cost function over the compound space of 6D kinematic rotations.[1]

2.1 3D Normalizing Flow-based Representations

Existing Work on 3D Human Priors. The method of [2] builds a density model to favor more probable poses over improbable ones. They use a mixture with 8 Gaussian modes $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, fitted to 1 million CMU poses. During optimization, the prior is evaluated to produce the log-likelihood of the pose. For numerical stability and to avoid excessive averaging effects, an approximation based on choosing the closest mode is used, which is not smooth, and may still lead to instability during mode switching.

For neural network models, [12] proposed a factorized adversarial network to learn the admissible rotation manifold of 3D poses, by relying on $N_j + 1$ discriminators, one for each joint, and one for the whole pose. The rotation limits for

[1] We have also considered quaternions, but our experiments showed these to be inferior even to angle-axis (AA), by at least 10%.

each joint are expected to be learned implicitly by each of the N_j discriminators, while the last one measures the probability of the combined pose. Learning rotation matrices (as opposed to angle-axis based) discriminators, is beneficial in avoiding the non-continuous nature of the angle-axis representation, but trades off increasing representational redundancy and consequently dimensionality.

Another approach has been pursued by [30], where the authors use a variational auto-encoder for 3D poses. The reconstruction loss is the mean per-vertex error between the input posed mesh and the reconstructed one. The latent representation can be used as a prior, by querying the log-likelihood of a given pose. Our experiments with VAEs constructed on top of kinematic representations (joint angles, rotations) showed that those have poor performance compared to our proposed models. The more sophisticated approaches used in VPoser [30] rely on losses defined on meshes rather than kinematics, but meshes inevitably introduce artefacts due to e.g. imperfect skinning. Moreover, VAEs need to balance two terms – the reconstruction loss and a KL divergence, which leads to a compromise: either the latent space is not close to Gaussian or/and decoding is imperfect. Our normalizing flow approach ensures that reconstruction loss is perfect (by the bijectivity of NFlow's construction) and during training we only optimize against the simpler Gaussian latent space objective.

Normalizing Flow Priors. In this paper we propose different normalizing flow-based prior representations, to our knowledge used for the first time in modeling 3D human pose. A normalizing flow [34, 4, 5, 16] is a sequence of invertible transformations applied to the original distribution. The end-result is a warped (latent) space with a potentially simple and tractable density function, e.g. $\mathbf{z} \sim \mathcal{N}(0; \mathbf{I})$). We consider $\boldsymbol{\theta} \sim p^*(\boldsymbol{\theta})$ sampled from an unknown distribution. One way to learn it is to use a dataset $\mathcal{D}$ (e.g. from CMU or Human3.6M) and maximize data log-likelihood with respect to a parametric model $p_{\boldsymbol{\phi}}(\boldsymbol{\theta})$

$$\max_{\boldsymbol{\phi}} \sum_{\boldsymbol{\theta} \in \mathcal{D}} \log p_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\phi}$ are the parameters of the generative model. If we choose $\mathbf{z} = \mathbf{f}_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ where $\mathbf{f}_{\boldsymbol{\phi}}$ is a component-wise invertible transformation, one can rewrite the log-probability under a change of variables

$$\log p_{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \log p_{\boldsymbol{\phi}}(\mathbf{z}) + \log \left| \det(d\mathbf{z}/d\boldsymbol{\theta}) \right| \tag{2}$$

Dropping the subscript $\boldsymbol{\phi}$, if $\mathbf{f}$ is the composition of multiple bijections $\mathbf{f}_i$, with intermediate output $\mathbf{h}_i$, (2) becomes

$$\log p_{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \log p_{\boldsymbol{\phi}}(\mathbf{z}) + \sum_{i=1}^{K} \log \left| \det(d\mathbf{h}_i/d\mathbf{h}_{i-1}) \right| \tag{3}$$

where $\mathbf{h}_0 = \boldsymbol{\theta}$ and $\mathbf{h}_K = \mathbf{z}$, and $p_{\boldsymbol{\phi}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$ is chosen as a spherical multivariate Gaussian distribution. State-of-the-art flow architectures are based on auto-regressive versions, such as the Masked Autoregressive Flow (MAF) [29], Inverse Autoregressive Flow (IAF) [17], NICE [4], MADE [7] or Real-NVP

[5]. In our experiments, we found MAF/IAF/MADE to be too slow given our representation and dataset size, with no measurable improvement over a Real-NVP. A Real-NVP step takes as input a variable $\mathbf{x}$ and outputs the transformed variable $\mathbf{y}$, under the following rules

$$\mathbf{y}_{1:d} = \mathbf{x}_{1:d}, \quad \mathbf{y}_{d+1:D} = \mathbf{x}_{d+1:D} \odot \exp \mathbf{s} + \mathbf{t}, \tag{4}$$

where $\mathbf{s}$ and $\mathbf{t}$ are shift-and-scale vectors that can be modelled as neural network outputs, i.e. $(\mathbf{s}, \mathbf{t}) = \text{NN}(\mathbf{x}_{1:d})$, and d is the splitting location of the current D-dimensional variable. The '$\odot$' operator represents the pointwise product, while 'exp' is the exponential function. In order to chain multiple Real-NVP steps, one has to ensure that order is not constant, otherwise the first d-dimensions would not be transformed. Typically, $\mathbf{x}$ is permuted before the operation. Because, in our case, $\boldsymbol{\theta}$ has moderate although sufficiently large size, we introduce a trainable, fully-connected layer before each NVP step. This is fast and results in better models. We also experiment with a lower capacity model, which replaces the Real-NVP with a simple parametric ReLU, as activation function. We do not use batch normalization. We found that we can trade a bit of accuracy (given by RealNVP) for a standard MLP that is faster and requires less memory. For the same network depth, the Real-NVP variant had 2x the number of parameters, and had marginal performance benefits (2%). More details can be found in the Sup. Mat.

For optimization-based inference or neural network training, we can parameterize the problem either in the latent (warped) space, or in the ambient (original) kinematic space, given the exact connection between them. Our empirical studies show that directly predicting (or optimizing) the latent representation always yields better results over working in the ambient space (see table 1).

In fig. 1 we show a sample pose interpolation in latent space.

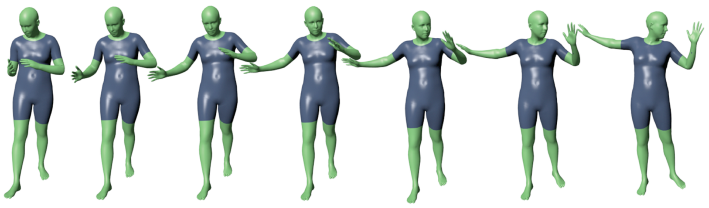

Fig. 1. From left to right: interpolation in latent space for normalizing flow, for two (begin and end) normal random codes. Notice smooth results, plausible human poses.

Optimization. To optimize normalizing flow representations, we assume normalization variable $\boldsymbol{\theta}$, Gaussian variable $\mathbf{z} = \mathbf{f}(\boldsymbol{\theta})$, and $\boldsymbol{\theta} = \mathbf{f}^{-1}(\mathbf{z})$, given that $\mathbf{f}$ is bijective. We define the normalizing flow prior as the negative log-likelihood in ambient $\psi_{nf}(\mathbf{f}(\boldsymbol{\theta})) = -\log p_{\boldsymbol{\phi}}(\mathbf{f}(\boldsymbol{\theta}))$ or, equivalently in latent space, $\psi_{nf}(\mathbf{z}) = -\log p_{\boldsymbol{\phi}}(\mathbf{z})$. Then, for any objective function or loss defined as $L(\boldsymbol{\theta}, \boldsymbol{\beta})$,

we have either the option of working (i.e. predicting or optimizing) in the **ambient space** and back-projecting in the latent space at each step

$$\arg\min_{\boldsymbol{\theta},\boldsymbol{\beta}} L(\boldsymbol{\theta},\boldsymbol{\beta}) + \psi_{nf}(\mathbf{f}(\boldsymbol{\theta})) \tag{5}$$

or the option to operate in the **latent space** directly

$$\arg\min_{\mathbf{z},\boldsymbol{\beta}} L(\mathbf{f}^{-1}(\mathbf{z}),\boldsymbol{\beta}) + \psi_{nf}(\mathbf{z}) \tag{6}$$

Both approaches are differentiable and we will evaluate them in §3.

2.2 Differentiable Semantic Alignment Loss

In order to be able to efficiently learn using weak supervision (e.g. just images of people), one needs a measure of prediction quality during the different phases of model training. In this work we explore forms of structured feedback by considering detailed correspondences between the different body part vertices of our 3D human body mesh (projected in the image), and the semantic human body part segmentation produced by another neural network.

As presented by [45], an Iterated Closest Point (ICP)-style cost for body part alignment can be designed in 2D (for 3D this is quite common e.g. [46]). Given a set of N_b body parts, their semantic image segmentation $\{\mathbf{S}_i \subset \mathbb{R}^2\}$ and associated mesh vertices of similar type $\{\mathbf{M}(\mathbf{V},k) \subset \mathbb{R}^3\}$ (i.e. the 3D vertex set of body part k), a distance between the set of semantic segmentation regions and the 3D mesh vertex projections (using an operator Π) can be defined as the first term of (7). This term encourages pixels of a particular semantic body type (e.g. torso, head or left lower arm) to attract projected model vertices with the same body part label. Depending on the sizes of the image regions with particular labels, and the corresponding number of vertices, the minimum of this function is not necessarily achieved only when all vertices are inside the body part. Consequently, we add a complementary loss, encouraging good overlap between model projections and image regions of corresponding semantics

$$L_{BA}(\mathbf{S},\mathbf{V}) = \sum_{k=1}^{N_b} \sum_{\mathbf{p}\in\mathbf{S}_k} \min_{\mathbf{v}\in\mathbf{M}(\mathbf{V},k)} \|\mathbf{p}-\Pi(\mathbf{v})\| + \sum_{k=1}^{N_b} \sum_{\mathbf{v}\in\mathbf{M}(\mathbf{V},k)} \min_{\mathbf{p}\in\mathbf{S}_k} \|\mathbf{p}-\Pi(\mathbf{v})\| \tag{7}$$

We will refer to the two terms as the forward semantic segmentation loss and the backward loss, respectively. Compared to state-of-the-art differentiable rendering techniques like [15], this loss has exact gradients, because we express it as an explicit objective connecting semantic image masks and mesh vertex projections. Furthermore, our method is designed for categorical masks and only defined for regions explained by the vertex projections of our model, rather than all the image pixels. The process is naturally parallelizable, and we offer a GPU implementation.

2.3 Network Architecture

Our architecture is based on a multistage deep convolutional neural network to predict human body joints, semantic segmentation of body parts, as well as 3D body pose and shape. The network consists of multiple modules, and has multiple losses, each corresponding to a different prediction task, but it can be run with a subset of the losses under different levels of supervision ranging from full to none. The first module takes as input the image and outputs keypoint (body joint) heatmaps [3]. We extract the joint positions from the heatmaps and obtain $\mathbf{J}_{2d} = \{\mathbf{J}_i, i = 1 \dots N_j\}$. The next module computes semantic body part segmentations by processing images and the keypoint heatmaps obtained by the keypoint prediction module. The outputs are semantic segmentation heatmaps for each body part (see fig. 2), $\mathbf{S} = \{\mathbf{S}_i, i = 1 \dots N_b\}$. The last module predicts pose and shape parameters. It takes as input the outputs from previous modules and produces $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$. For the camera, we adopt a perspective projection model. We fix the intrinsics and estimate translation by means of fitting the predicted 3D skeleton to 2D joint detections (that step alone requires a weak perspective approximation, see Sup. Mat.).

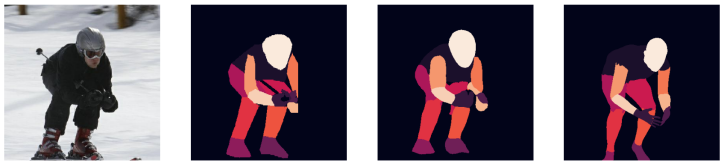

Fig. 2. From left to right: Original image, ground truth semantic body part segmentation mask from MSCOCO 2014, predicted segmentation mask, projected semantic mask of our 3D mesh.

3D Pose and Shape. The goal of the 3D pose layers is to predict the pose and shape parameters $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$. The associated network is similar to the ones of the previous two modules. A stack of convolutional stages is created with losses on each stage to reinforce the weights and avoid vanishing gradients [41, 3]. The architecture of each 3D regressor stage is composed of a stack of 5 x 2D convolutional layers with 128 feature maps, 7x7 kernels, `relu` activations, followed by another 2D convolutional module with 128 layers and 1x1 kernels. The last layer is a 2D convolutional layer, has no activation function and the number of heatmaps is equal to the number of predicted parameters. Two separate dense layers are used to output $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$.

Supervised and Weakly Supervised Losses We train our network by using a combination of fully and weakly supervised losses. The fully supervised training regime assumes complete ground truth on pose, shape. A predicted posed mesh $\mathbf{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ with N_v associated vertices $\mathbf{V} = \{\mathbf{v}_i, i = 1 \dots N_v\}$ has ground truth

$\mathbf{M}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$ with vertices $\{\widehat{\mathbf{v}}_i\}$. We define the following MSE losses, respectively, on the mesh

$$L_V = \frac{1}{N_v} \sum_{i=1}^{N_v} \|\mathbf{v}_i - \widehat{\mathbf{v}}_i\|_2^2 \tag{8}$$

pose and shape parameters

$$L_{\boldsymbol{\theta}} = \frac{1}{N_j} \sum_{i=1}^{N_j} \left\|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i\right\|_2^2, \quad L_{\boldsymbol{\beta}} = \frac{1}{N_s} \sum_{i=1}^{N_s} \left\|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i\right\|_2^2 \tag{9}$$

The supervised loss combines previously defined losses

$$L_{fs} = L_V + L_{\boldsymbol{\theta}} + L_{\boldsymbol{\beta}} \tag{10}$$

For the weakly supervised case, the predicted mesh $\mathbf{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ is projected into the image. Denote the projected skeleton joints by $\mathbf{J}_{2d} = \{\mathbf{J}_i\}$, the estimated (or ground truth) 2D joint positions by $\widehat{\mathbf{J}}_{2d} = \{\widehat{\mathbf{J}}_i\}$, and the semantic body part segmentation maps by $\widehat{\mathbf{S}} = \{\widehat{\mathbf{S}}_i, i = 1 \dots N_b\}$. The weakly supervised regime assumes access to large 3D mocap datasets, e.g. CMU – in order to construct kinematic priors – but without the corresponding images. Additionally we also rely on images in the wild, with only 2D body joint or semantic segmentation maps ground truth. Our weakly-supervised model relies on all practically useful data in order to bootstrap a self-supervised system at later stages. Hence we do not discard 3D data when we have it, and aim to use it to circumvent the missing link: images in the wild with 3D pose and shape ground truth. Then, one can define weakly supervised losses for *keypoint alignment:* $L_{KA} = \frac{1}{N_j} \sum_{i=1}^{N_j} \left\|\mathbf{J}_i - \widehat{\mathbf{J}}_i\right\|_2^2$, *semantic body-part alignment:* $L_{BA}\left(\widehat{\mathbf{S}}, \mathbf{V}\right)$, and *the prior:* $L_\psi = \psi_{nf}\left(\mathbf{f}(\boldsymbol{\theta})\right)$ (or $\psi_{nf}\left(\mathbf{z}\right)$ when working in the latent space). The weakly supervised loss is a combination of multiple losses, plus a term that regularizes the shape parameters

$$L_{ws} = L_{KA} + L_{BA} + L_\psi + \|\boldsymbol{\beta}\|_2^2 \tag{11}$$

The total loss will be $L_{total} = L_{fs} + L_{ws}$. For a graceful transition between supervision regimes, during fully supervised training we use L_{total}, then switch to L_{ws} in the weakly supervised phase.

3 Experiments

Datasets. We run our fully supervised experiments on the Human80K (H80K) – a representative subset sampled from Human3.6M (H3.6M) [8]. We also use H80K in order to train pose priors and for optimization experiments. We report errors in the form of MPJPE (mean per-joint position error) and MPVPE (mean per-vertex position error) all in 3D.

We split the training set of H80K (composed of $\approx 54,000$ images) into train, eval and test. As there are no publicly available statistical body model fittings

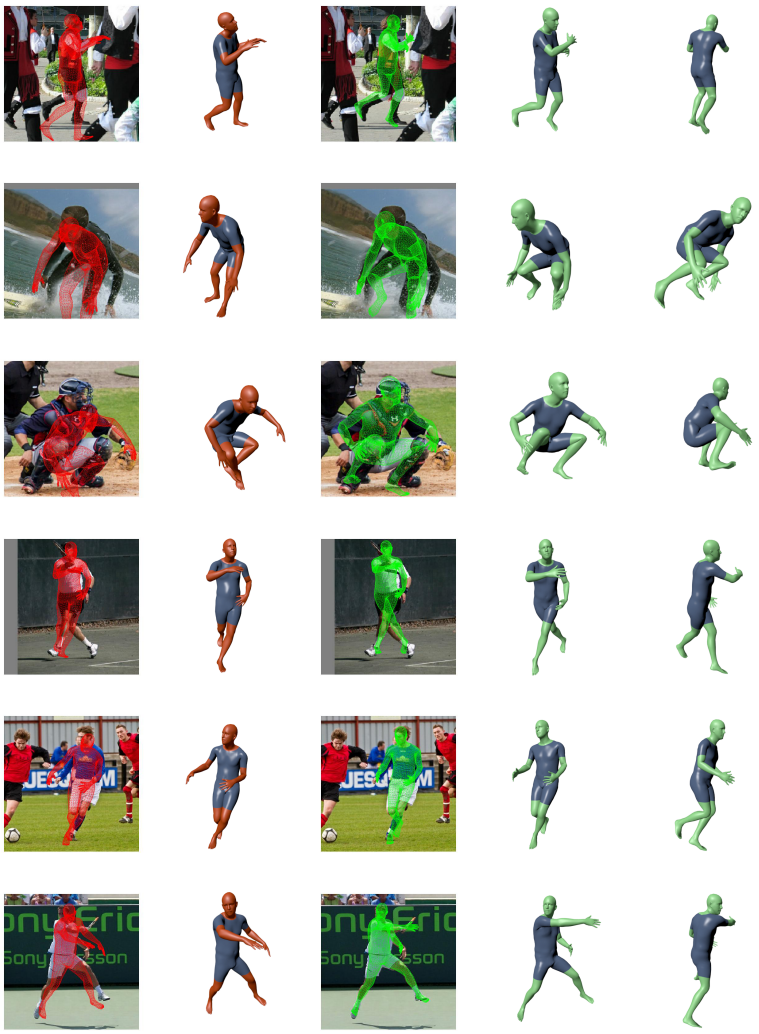

Fig. 3. Reconstruction results of models trained weakly-supervised using COCO (best seen in color). Starting from a network fully supervised on H80K (**red**), we fine-tune with a weakly-supervised loss (**green**) and a normalizing flow kinematic prior. Notice considerable improvement in both alignment and the perceptual 3D estimates. Last column shows a different view angle for the WS estimate.

for H80K data, we had to build them ourselves. Based on the ground truth 3D joint positions $\widehat{\mathbf{J}}_{3D}$ (this is used to retrieve pose $\widehat{\boldsymbol{\theta}}$) and the available 3D subject scans (used to retrieve shape $\widehat{\boldsymbol{\beta}}$) provided with the dataset, we optimize a fitting objective (solved using BFGS). We then project the 3D meshes associated to motion captured body configurations in each frame, to obtain ground truth 2D annotations, $\widehat{\mathbf{J}}_{2d}$ and $\widehat{\mathbf{S}}_{2d}$. We thus have full supervision on H80K in the form of $\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{J}}_{2d}, \widehat{\mathbf{S}}_{2d}\right)$ for each image in the training set.

We also use CMU [11] and AMASS [24] to train 3D pose priors. Both datasets have publicly available kinematic model fittings and we used the $\widehat{\boldsymbol{\theta}}$ values to train our normalizing pose model $p_{\boldsymbol{\phi}}(\boldsymbol{\theta})$. This results in priors over ambient and latent spaces, $\psi_{nf}(\mathbf{f}(\boldsymbol{\theta}))$ and $\psi_{nf}(\mathbf{z})$, respectively. Similar models were trained for H80K.

For weakly supervised learning 'in the wild', we use a subset of 15,000 images from COCO 2014 [21]. The dataset has no 3D ground truth, but offers 2D annotations for human body joints $\widehat{\mathbf{J}}_{2d}$, as well semantic segmentation of body parts $\widehat{\mathbf{S}}_{2d}$. We split the data in 14,000 examples for training and 1,000 for testing, and use it for building the weakly supervised models. We refer to models trained using 2D body joints and semantic body part losses as KA and BA, respectively.

Optimization with Different Priors and Losses. In order to analyze the impact of priors and semantic segmentation losses on optimization, we choose H80K where ground-truth is available for all components including 3D camera, pose and shape. We perform non-linear optimization with the objective function as defined in (11), where L_{ψ} is changed to accommodate all the various priors, and L_{KA} and L_{BA} are studied both together and independently.

We evaluate different prior types: **i)** $\psi_{gmm}(\boldsymbol{\theta})$ – GMM [2], **ii)** $\psi_{nf}(\mathbf{f}(\boldsymbol{\theta}))$ and $\psi_{nf}(\mathbf{z})$ – normalizing flow in ambient and latent space, as given by (5), and (6), using either the angle-axis or the 6D representation, **iii)** $\psi_{VPoser}(\mathbf{z})$ – the variational auto-encoder VPoser of [30], **iv)** $\psi_{hmr}(\boldsymbol{\theta})$ – the discriminator of [12].

We also evaluate different datasets (CMU, H80K, AMASS) for prior construction, different loss functions based on either body joints/keypoints or semantic segmentation of body parts (KA and BA). To directly compare with VPoser, we train a light-weight normalizing flow prior ($\approx 93,000$ parameters compared to $\approx 344,000$ for VPoser), with the same operating speed, and constructed on the same dataset (AMASS) and train/test splits.

To isolate confounding factors, optimization is performed using the ground-truth 2D joints (KA) and body part labels (BA), under a perspective projection model, by using the loss defined at (11). Optimization relies on BFGS with analytical Jacobians, obtained through automatic differentiation. We start with four different initializations and report the solution with the smaller loss (N.B. this does not require observing the ground truth). We consider four different global rotations, and initialize parameters with $\mathbf{0}$, for pose (either in ambient or latent space) and shape.

We test the model on 500 images and report results in table 1. The best results are achieved by normalizing flow priors when optimization is performed in latent space. By using both keypoint and body part alignment-based self-supervision,

the results improve. The 6D rotation representation has a slight edge over the angle-axis. The light-weight normalizing flow trained on AMASS is the best performer, surpassing VPoser even at a third of its capacity. Note that we do not have to balance two terms (the reconstruction loss and the KL-divergence), as normalizing flows support exact latent-variable inference. Additionally, VPoser requires posing meshes *during training*, whereas normalizing flow models do not.

Method *prior, dataset, representation, features*	**Error (cm)** MPJPE/MPVPE
$\psi_{gmm}(\boldsymbol{\theta})$, CMU, AA, KA	7.9/10.4
$\psi_{gmm}(\boldsymbol{\theta})$, CMU, AA, KA + BA	6.9/9.6
$\psi_{VPoser}(\mathbf{z})$, AMASS, 6D, KA	4.6/6.7
$\psi_{nf}(\mathbf{z})$, AMASS, 6D, KA	**4.3/6.0**
$\psi_{hmr}(\boldsymbol{\theta})$, H3.6M, RM, KA	11.9/15.3
$\psi_{nf}(\mathbf{f}(\boldsymbol{\theta}))$, CMU, AA, KA	6.2/8.4
$\psi_{nf}(\mathbf{f}(\boldsymbol{\theta}))$, CMU, AA, KA + BA	6.0/8.1
$\psi_{nf}(\mathbf{z})$, CMU, AA, KA	5.0/7.1
$\psi_{nf}(\mathbf{z})$, CMU, AA, KA + BA	4.9/6.9
$\psi_{nf}(\mathbf{f}(\boldsymbol{\theta}))$, CMU, 6D, KA	6.1/8.4
$\psi_{nf}(\mathbf{f}(\boldsymbol{\theta}))$, CMU, 6D, KA + BA	5.8/8.0
$\psi_{nf}(\mathbf{z})$, CMU, 6D, KA	5.1/6.8
$\psi_{nf}(\mathbf{z})$, CMU, 6D, KA + BA	4.8/6.6
$\psi_{nf}(\mathbf{f}(\boldsymbol{\theta}))$, H80K, AA, KA	5.4/7.5
$\psi_{nf}(\mathbf{z})$, H80K, AA, KA	4.4/6.1

Table 1. Optimization-based pose and shape estimation experiments with evaluation on the ground truth of H80K dataset. Priors are learned on the training sets of CMU, AMASS or H80K. The HMR discriminator has the largest errors, as it was arguably designed for use with deep neural network losses, and not for model fitting. Optimizing in latent space (using normalizing flows) and semantic alignment always helps. The 6D representation performs slightly better than angle-axis. The best performers are objective functions that include normalizing flow priors trained on H80K or AMASS. VPoser performs slightly worse than our normalizing flow prior, even though it also encodes and decodes 6D rotations. *Notation:* AA = angle-axis representation, 6D = 6 dimensions rotation representation, RM = rotation matrices, KA = keypoint alignment, BA = body alignment.

Fully to Weakly Supervised Transfer Learning. We present experiments and ablation studies showing how the weakly supervised training of shape and pose parameters $(\boldsymbol{\theta}, \boldsymbol{\beta})$ can be successful in conjunction with the proposed normalizing flow priors and self-supervised losses.

Percentage Supervised	0%	20%	40%	60%	80%	100%
FS (mm)	649/677	117/136	101/118	93/109	86/102	83/97.15
WS (mm)	**123/140**	**97/111**	**92/108**	**90/106**	**85/101**	84/98.85

Table 2. Ablations on H80K, reported as MPJPE/MPVPE metrics in millimeters. Notice the impact of weakly supervised losses (WS), especially in the fully supervised (FS) regime with small training sets, as well as for the model initialized randomly (column two, 0% supervision).

For this study, we split H80K into two parts where we keep 5 subjects for training (S1, S5, S6, S7 and S8) and two subjects (S9, S11) for testing.

We further split the training set into partitions of 20%, 40%, 60%, 80%, 100%. We initially train the network fully supervised (FS) on the specific partition of the data using L_{fs} loss. We train the fully supervised model for 30 epochs, then continue in a weakly supervised (WS) regime based on L_{ws} on all the data. In table 2 we report MPJPE/MPVPE for the ablation study. Notice that in all cases weak supervision improves performance whenever additional image data is available.

We also check that our methodology compares favorably to a similar method HMR [12] which we retrained on H80K. In this case our model achieves 84mm MPJE whereas HMR has 88mm.[2] We were not able to train on their split and retargeting of H3.6M, as their training data was not available.

Weakly Supervised Transfer for Images in the Wild. In order to validate our network predictions beyond a motion capture laboratory, 'in the wild', we refined the network on the subset of COCO which has body part labelling available. We started with a network pre-trained on H80K, then continued training on COCO using the complete loss. As ground truth 3D is not available for COCO, we monitor errors between ground truth and estimated 2D projections of the 3D model joints, and the IoU semantic body part alignment metrics. As shown in fig. 4, in all cases the pixel error of the projected 2D joints decreased consistently, as a result of weakly supervised fine tuning. A similar trend can be seen for the IoU metric computed for body part alignment, illustrating the importance of a segmentation loss. We explicitly run two configurations, one in which we only use the keypoints alignment (KA) and another based on body part alignment (KA+BA).

A potentially interesting question is whether the 3D prediction is affected by a self-supervised refinement. We run experiments on 3DPW [25] which consists of $\approx 60,000$ images containing one or more humans performing various actions in the wild. The subjects were recorded using IMUs so shape and pose parameters were recovered. We used the training data as supervision, and evaluate on the test set. We report results for a model trained only with full supervision, as well as results of refining the fully supervised (feed-forward) estimate by fur-

[2] Based on HMR's Github repository, we identify a total of ≈27M trainable parameters. Our model has 6 stages, each with $5 \times 7 \times 7 \times 128 \times 128$ parameters resulting in ≈24M trainable parameters.

	Method	MPJPE (mm)	MPJPE-PA (mm)
STATIC	HMR [12]	-	81.3
	Kanazawa et al. [14]	-	72.6
	SPIN [19] (static fits)	-	66.3
	SPIN [19] (best)	-	59.2
	FS	95	61.3
	FS+OPT (KA)	95	60.3
	FS+OPT (KA+BA)	91.4	58.87
	FS+WS (KA+BA)	**90.0**	**57.1**
VIDEO	VIBE [18](16 frames)	82.9	51.9
	FS+OPT(KA+BA+S, 16 frames)	82.8	52.2
	FS+WS+OPT(KA+BA+S, 4 frames)	84.5	54.5
	FS+WS+OPT(KA+BA+S, 8 frames)	82.0	51.4
	FS+WS+OPT(KA+BA+S, 16 frames)	**80.2**	**49.8**

Table 3. Results on the 3DPW test set for two regimes: **static** and **video**. FS is fully supervised, FS+OPT are predictions from FS with optimization. FS+WS are results for self-supervised refinement of the FS model on MS COCO. 'S' stands for smoothing in the video regime. MPJPE is the mean per joint position error, whereas MPJPE-PA is the error after Procrustes alignment. **Static**: we observe that the self-supervised training did not affect the performance of the 3D predictions. The semantic alignment loss reduces error more than only keypoints alignment. Perceptually, image alignment is also much better for BA than KA, even when it does not immediately produce significant 3D quantitative improvements. **Video**: the best performer is our FS+WS (KA+BA) model, further optimized over 16 frames with the temporal smoothing term.

ther optimizing the KA, and KA+BA losses against the predicted 2D outputs (keypoint and body part alignment). After training the network in the weakly supervised regime we obtain better accuracy, showing that 3D prediction quality is preserved. We show the results in table 3. To the best of our knowledge, these are the lowest errors reported so far on the 3DPW test set in a static setting.

Temporal optimization. We also experiment in the temporal setting, on batches of 4, 8 and 16 consecutive frames drawn from the 3DPW dataset. Starting from the best results obtained per frame in the static setting, we do a whole batch optimization. Different from the L_{ws} objective, now the shape parameters are tied across frames, with an additional term that enforces smoothness between adjacent temporal pose parameters (in latent space):

$$L_{smooth} = \sum_{t=2}^{N_f} \left\| \mathbf{z}^t - \mathbf{z}^{t-1} \right\|_2^2 \tag{12}$$

The weight for this term is set to be 50× the weight of the prior, as we expect a lower variance for pose dynamics. We compare our method with the recent work of [18], showing the results in table 3. As in the static setting, these are also the lowest errors reported so far.

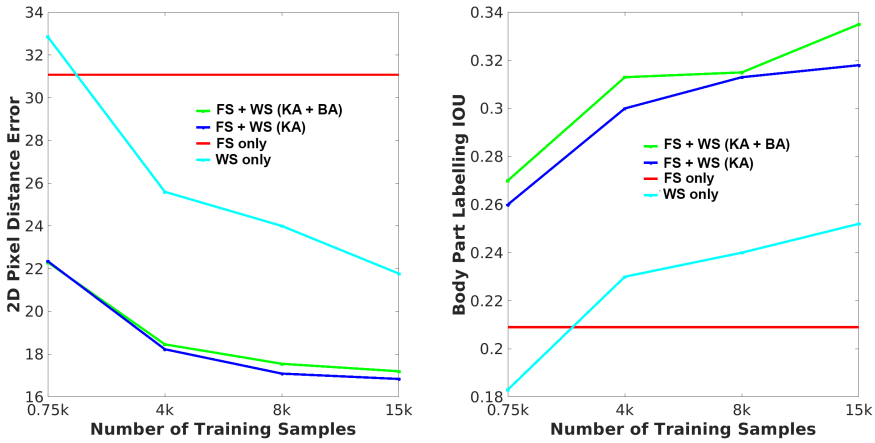

Fig. 4. *Left and Right*: Weakly supervised experiments on COCO with different loss combinations (KA, KA+BA) and different amounts of training data. The baseline is obtained by running the network trained fully supervised on H80K. WS Only is trained only on COCO.

4 Conclusions

We have presented large scale weakly supervised deep learning-based models for 3D human pose and shape estimation from monocular images and video. Key to scalability is unlocking the ability to exploit human statistics implicitly available in large, diverse image repositories, which however do not come with detailed 3D pose or shape supervision. Key to making such approaches feasible, in terms of identifying model parameters with good generalization performance, is the ability to design training losses that are tightly controlled by both the existing prior knowledge on human pose and shape, and by the image and video evidence.

We introduce latent normalizing flow representations and dynamical models, as well as fully differentiable, structured, semantic body party alignment (re-projection) loss functions which provide informative feedback for self-supervised learning. In extensive, large-scale experiments, using both motion capture datasets like CMU, Human3.6M, AMASS, or 3DPW, as well as 'in the wild' repositories like COCO, we show that our proposed methodology achieves state-of-the-art results in both images and video, supporting the claim that constructing accurate models based on large-scale weak supervision 'in the wild' is possible.

References

1. Arnab, A., Doersch, C., Zisserman, A.: Exploiting temporal context for 3d human pose estimation in the wild (2019)
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In: ECCV (2016)
3. Cao, Z., Simon, T., Wei, S., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
4. Dinh, L., Krueger, D., Bengio, Y.: Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516 (2014)
5. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. arXiv preprint arXiv:1605.08803 (2016)
6. Doersch, C., Zisserman, A.: Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In: Advances in Neural Information Processing Systems. pp. 12929–12941 (2019)
7. Germain, M., Gregor, K., Murray, I., Larochelle, H.: Made: Masked autoencoder for distribution estimation. In: ICML (2015)
8. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. PAMI (2014)
9. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: International Conference on Computer Vision (ICCV) (2019)
10. Jackson, A.S., Manafas, C., Tzimiropoulos, G.: 3d human body reconstruction from a single image via volumetric regression. In: ECCV (2018)
11. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: CVPR (2018)
12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
13. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5614–5623 (2019)
14. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Computer Vision and Pattern Recognition (CVPR) (2019)
15. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: CVPR (2018)
16. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: NeurIPS (2018)
17. Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improved variational inference with inverse autoregressive flow. In: NeurIPS (2016)
18. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. arXiv preprint arXiv:1912.05656 (2019)
19. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2252–2261 (2019)
20. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4501–4510 (2019)
21. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014), http://arxiv.org/abs/1405.0312

22. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. SIGGRAPH (2015)
23. Loper, M.M., Black, M.J.: Opendr: An approximate differentiable renderer. In: ECCV (2014)
24. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. arXiv preprint arXiv:1904.03278 (2019)
25. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (sep 2018)
26. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV (2017)
27. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) (2017)
28. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P.V., Schiele, B.: Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In: 3DV (2018)
29. Papamakarios, G., Pavlakou, T., Murray, I.: Masked autoregressive flow for density estimation. In: NeurIPS (2017)
30. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A., Tzionas, D., Black, M.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
31. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: CVPR (2017)
32. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: CVPR (2018)
33. Popa, A., Zanfir, M., Sminchisescu, C.: Deep Multitask Architecture for Integrated 2D and 3D Human Sensing. In: CVPR (2017)
34. Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770 (2015)
35. Rhodin, H., Robertini, N., Richardt, C., Seidel, H.P., Theobalt, C.: A versatile scene model with differentiable visibility applied to generative pose estimation. In: ICCV (2015)
36. Rogez, G., Schmid, C.: Mocap-guided data augmentation for 3d pose estimation in the wild. In: NeurIPS (2016)
37. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5349–5358 (2019)
38. Tekin, B., Marquez Neila, P., Salzmann, M., Fua, P.: Learning to fuse 2d and 3d image cues for monocular body pose estimation. In: ICCV (2017)
39. Tung, H.Y., Tung, H.W., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture. In: NeurIPS (2017)
40. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: BodyNet: Volumetric inference of 3D human body shapes. In: ECCV (2018)
41. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)
42. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6184–6193 (2020)

43. Xu, Y., Zhu, S.C., Tung, T.: Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7760–7770 (2019)
44. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3d human pose estimation in the wild by adversarial learning. In: CVPR (2018)
45. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: CVPR (2018)
46. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In: CVPR (2017)
47. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: ICCV (2017)
48. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. arXiv preprint arXiv:1812.07035 (2018)