End-to-end Recovery of Human Shape and Pose

Angjoo Kanazawa[1], Michael J. Black[2], David W. Jacobs[3], Jitendra Malik[1]
[1]University of California, Berkeley
[2]MPI for Intelligent Systems, Tübingen, Germany, [3]University of Maryland, College Park
{kanazawa,malik}@eecs.berkeley.edu, black@tuebingen.mpg.de, djacobs@umiacs.umd.edu

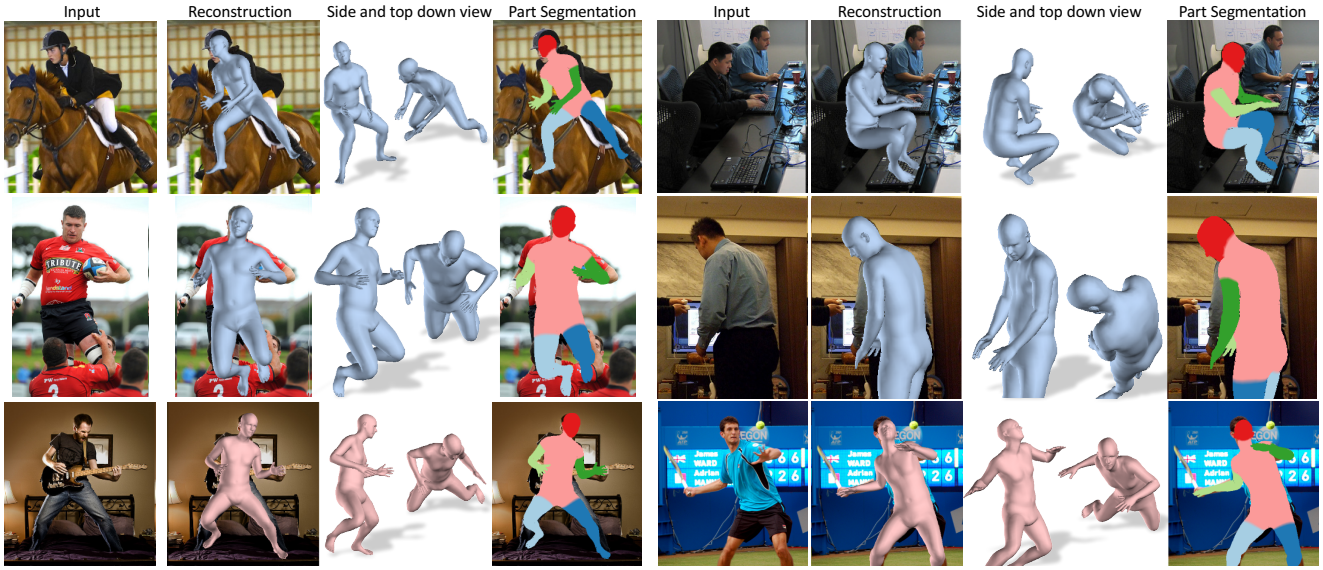

Figure 1: **Human Mesh Recovery (HMR): End-to-end adversarial learning of human pose and shape**. We describe a real time framework for recovering the 3D joint angles and shape of the body from a single RGB image. The first two rows show results from our model trained with some 2D-to-3D supervision, the bottom row shows results from a model that is trained in a fully weakly-supervised manner without using any paired 2D-to-3D supervision. We infer the full 3D body even in case of occlusions and truncations. Note that we capture head and limb orientations.

Abstract

We describe Human Mesh Recovery (HMR), an end-to-end framework for reconstructing a full 3D mesh of a human body from a single RGB image. In contrast to most current methods that compute 2D or 3D joint locations, we produce a richer and more useful mesh representation that is parameterized by shape and 3D joint angles. The main objective is to minimize the reprojection loss of keypoints, which allows our model to be trained using in-the-wild *images that only have ground truth 2D annotations. However, the reprojection loss alone is highly underconstrained. In this work we address this problem by introducing an adversary trained to tell whether human body shape and pose parameters are real or not using a large database of 3D human meshes. We show that HMR can be trained with and* without *using any paired 2D-to-3D supervision. We do not rely on intermediate 2D keypoint detections and infer 3D pose and shape parameters directly from image pixels. Our model runs in real-time given a bounding box containing the person. We demonstrate our approach on various images* in-the-wild *and out-perform previous optimization-based methods that output 3D meshes and show competitive results on tasks such as 3D joint location estimation and part segmentation.*

1. Introduction

We present an end-to-end framework for recovering a full 3D mesh of a human body from a single RGB image. We use the generative human body model, SMPL [24], which parameterizes the mesh by 3D joint angles and a low-

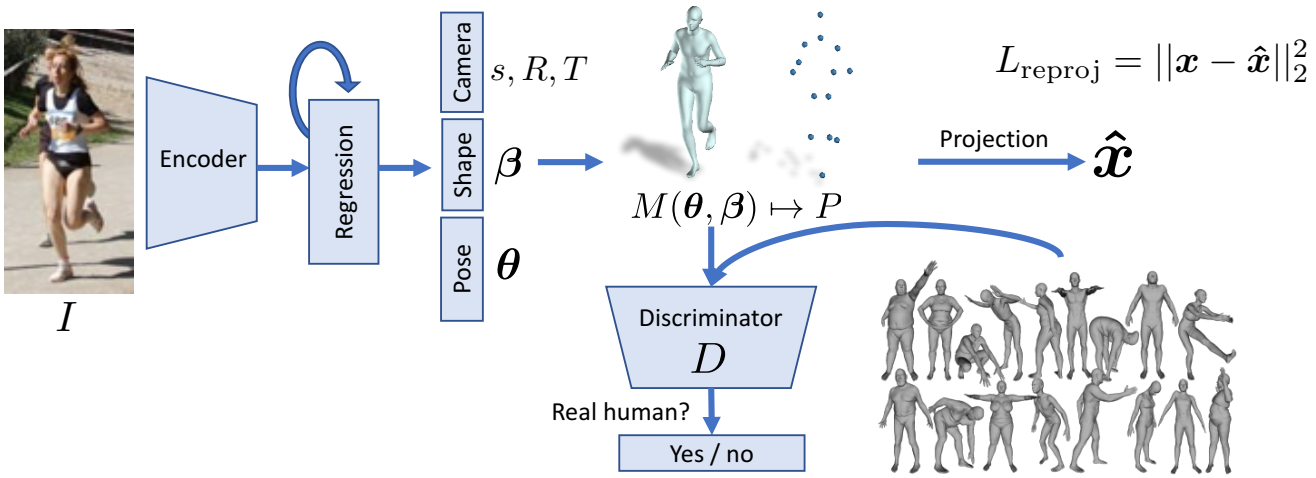

Figure 2: **Overview of the proposed framework.** An image I is passed through a convolutional encoder. This is sent to an iterative 3D regression module that infers the latent 3D representation of the human that minimizes the joint reprojection error. The 3D parameters are also sent to the discriminator D, whose goal is to tell if these parameters come from a real human shape and pose.

dimensional linear shape space. As illustrated in Figure 1, estimating a 3D mesh opens the door to a wide range of applications such as foreground and part segmentation, which is beyond what is practical with a simple skeleton. The output mesh can be immediately used by animators, modified, measured, manipulated and retargeted. Our output is also holistic – we always infer the full 3D body even in cases of occlusion and truncation.

Note that there is a great deal of work on the 3D analysis of humans from a single image. Most approaches, however, focus on recovering 3D joint locations. We argue that these joints alone are not the full story. Joints are sparse, whereas the human body is defined by a surface in 3D space.

Additionally, joint locations alone do not constrain the full DoF at each joint. This means that it is non-trivial to estimate the full pose of the body from only the 3D joint locations. In contrast, we output the relative 3D rotation matrices for each joint in the kinematic tree, capturing information about 3D head and limb orientation. Predicting rotations also ensures that limbs are symmetric and of valid length. Our model implicitly learns the joint angle limits from datasets of 3D body models.

Existing methods for recovering 3D human mesh today focus on a multi-stage approach [5, 20]. First they estimate 2D joint locations and, from these, estimate the 3D model parameters. Such a stepwise approach is typically not optimal and here we propose an end-to-end solution to learn a mapping from image pixels directly to model parameters.

There are several challenges, however, in training such a model in an end-to-end manner. First is the lack of large-scale ground truth 3D annotation for *in-the-wild* images. Existing datasets with accurate 3D annotations are captured in constrained environments. Models trained on these datasets do not generalize well to the richness of images in the real world. Another challenge is in the inherent ambiguities in single-view 2D-to-3D mapping. Most well known is the problem of depth ambiguity where multiple 3D body configurations explain the same 2D projections [42]. Many of these configurations may not be anthropometrically reasonable, such as impossible joint angles or extremely skinny bodies. In addition, estimating the camera explicitly introduces an additional scale ambiguity between the size of the person and the camera distance.

In this paper we propose a novel approach to mesh reconstruction that addresses both of these challenges. A key insight is that there are large-scale 2D keypoint annotations of *in-the-wild* images and a separate large-scale dataset of 3D meshes of people with various poses and shapes. Our key contribution is to take advantage of these *unpaired* 2D keypoint annotations and 3D scans in a conditional *generative adversarial manner*. The idea is that, given an image, the network has to infer the 3D mesh parameters and the camera such that the 3D keypoints match the annotated 2D keypoints after projection. To deal with ambiguities, these parameters are sent to a discriminator network, whose task is to determine if the 3D parameters correspond to bodies of real humans or not. Hence the network is encouraged to output parameters on the human manifold and the discriminator acts as weak supervision. The network implicitly learns the angle limits for each joint and is discouraged from making people with unusual body shapes.

An additional challenge in predicting body model parameters is that regressing to rotation matrices is challenging. Most approaches formulate rotation estimation as a

classification problem by dividing the angles into bins [45]. However differentiating angle probabilities with respect to the reprojection loss is non-trivial and discretization sacrifices precision. Instead we propose to directly regress these values in an iterative manner with feedback. Our framework is illustrated in Figure 2.

Our approach is similar to 3D interpreter networks [33, 50] in the use of reprojection loss and the more recent adversarial inverse graphics networks [47] for the use of the adversarial prior. We go beyond the existing techniques in multiple ways:

1. We infer 3D mesh parameters directly from image features, while previous approaches infer them from 2D keypoints. This avoids the need for two stage training and also avoids throwing away valuable information in the image such as context.

2. Going beyond skeletons, we output meshes, which are more complex and more appropriate for many applications. Again, no additional inference step is needed.

3. Our framework is trained in an end-to-end manner. We out-perform previous approaches that output 3D meshes [5, 20] in terms of 3D joint error and run time.

4. We show results with and *without* paired 2D-to-3D data. Even without using any paired 2D-to-3D supervision, our approach produces reasonable 3D reconstructions. This is most exciting because it opens up possibilities for learning 3D from large amounts of 2D data.

Since there are no datasets for evaluating 3D mesh reconstructions of humans from in-the-wild images, we are bound to evaluate our approach on the standard 3D joint location estimation task. Our approach out performs previous methods that estimate SMPL parameters from 2D joints and is competitive with approaches that only output 3D skeletons. We also evaluate our approach on an auxiliary task of human part segmentation. We qualitatively evaluate our approach on challenging images in-the-wild and show results sampled at different error percentiles. Our model and code is available for research purposes at https://akanazawa.github.io/hmr/.

2. Related Work

3D Pose Estimation: Many papers formulate human pose estimation as the problem of locating the major 3D joints of the body from an image, a video sequence, either single-view or multi-view. We argue that this notion of "pose" is overly simplistic but it is the major paradigm in the field. The approaches are split into two categories: two-stage and direct estimation.

Two stage methods first predict 2D joint locations using 2D pose detectors [30, 49, 54] or ground truth 2D pose and then predict 3D joint locations from the 2D joints either by regression [26, 29] or model fitting, where a common approach exploits a learned dictionary of 3D skeletons [2, 34, 47, 37, 53, 54]. In order to constrain the inherent ambiguity in 2D-to-3D estimation, these methods use various priors [42]. Most methods make some assumption about the limb-length or proportions [4, 21, 32, 34]. Akhter and Black [2] learn a novel pose prior that captures pose-dependent joint angle limits. Two stage-methods have the benefit of being more robust to domain shift, but rely too much on 2D joint detections and may throw away image information in estimating 3D pose.

Video datasets with ground truth motion capture like HumanEva [38] and Human3.6M [16] define the problem in terms of 3D joint locations. They provide training data that lets the 3D joint estimation problem be formulated as a standard supervised learning problem. Thus, many recent methods estimate 3D joints directly from images in a deep learning framework [33, 43, 44, 51, 52]. Dominant approaches are fully-convolutional, except for the very recent method of Xiao *et al*. [40] that regresses bones and obtains excellent results on the 3D pose benchmarks. Many methods do not solve for the camera, but estimate the depth relative to root and use a predefined global scale based the average length of bones [33, 51, 52]. Recently Rogez *et al*. [36] combine human detection with 3D pose prediction. The main issue with these direct estimation methods is that images with accurate ground truth 3D annotations are captured in controlled MoCap environments. Models trained only on these images do not generalize well to the real world.

Weakly-supervised 3D: Recent work tackles this problem of the domain gap between MoCap and *in-the-wild* images in an end-to-end framework. Rogez and Schmid [35] artificially endow 3D annotations to images with 2D pose annotation using MoCap data. Several methods [27, 28, 51] train on both in-the-wild and MoCap datasets jointly. Still others [27, 28] use pre-trained 2D pose networks and also use 2D pose prediction as an auxiliary task. When 3D annotation is not available, Zhou *et al*. [51] gain weak supervision from a geometric constraint that encourages relative bone lengths to stay constant. In this work, we output 3D joint angles and 3D shape, which subsumes these constraints that the limbs should be symmetric. We employ a much stronger form of weak supervision by training an adversarial prior.

Methods that output more than 3D joints: There are multiple methods that fit a parametric body model to manually extracted silhouettes [8] and a few manually provided correspondences [12, 14]. More recent works attempt to automate this effort. Bogo *et al*. [5] propose SMPLify, an optimization-based method to recover SMPL parameters from 14 detected 2D joints that leverages multiple pri-

ors. However, due to the optimization steps the approach is not real-time, requiring 20-60 seconds per image. They also make *a priori* assumptions about the joint angle limits. Lassner *et al*. [20] take curated results from SMPLify to train 91 keypoint detectors corresponding to traditional body joints and points on the surface. They then optimize the SMPL model parameters to fit the keypoints similarly to [5]. They also propose a random forest regression approach to directly regress SMPL parameters, which reduces runtime at the cost of accuracy. Our approach out-performs both methods, directly infers SMPL parameters from images instead of detected 2D keypoints, and runs in real time.

VNect [28] fits a rigged skeleton model over time to estimated 2D and 3D joint locations. While they can recover 3D rotations of each joint after optimization, we directly output rotations from images as well as the surface vertices. Similarly Zhou *et al*. [52] directly regress joint rotations of a fixed kinematic tree. We output shape as well as the camera scale and out-perform their approach in 3D pose estimation.

There are other related methods that predict SMPL-related outputs: Varol *et al*. [48] use a synthetic dataset of rendered SMPL bodies to learn a fully convolutional model for depth and body part segmentation. DenseReg [13] similarly outputs a dense correspondence map for human bodies. Both are 2.5D projections of the underlying 3D body. In this work, we recover all SMPL parameters and the camera, from which all of these outputs can be obtained.

Kulkarni *et al*. [19] use a generative model of body shape and pose with a probabilistic programming framework to estimate body pose from single image. They deal with visually simple images and do not evaluate 3D pose accuracy. More recently Tan *et al*. [41] infer SMPL parameters by first learning a silhouette decoder of SMPL parameters using synthetic data, and then learning an image encoder with the decoder fixed to minimize the silhouette reprojection loss. However, the reliance on silhouettes limits their approach to frontal images and images of humans without any occlusion. Concurrently Tung *et al*. [46] predict SMPL parameters from an image and a set of 2D joint heatmaps. The model is pretrained on a synthetic dataset and fine-tuned at test time over two consecutive video frames to minimize the reprojection loss of keypoints, silhouettes and optical flow. Our approach can be trained without any paired supervision, does not require 2D joint heatmaps as an input and we test on images without fine-tuning. Additionally, we also demonstrate our approach on images of humans in-the-wild [22] with clutter and occlusion.

3. Model

We propose to reconstruct a full 3D mesh of a human body directly from a single RGB image I centered on a human in a feedforward manner. During training we assume that all images are annotated with ground truth 2D joints. We also consider the case in which some have 3D annotations as well. Additionally we assume that there is a pool of 3D meshes of human bodies of varying shape and pose. Since these meshes do not necessarily have a corresponding image, we refer to this data as *unpaired* [55].

Figure 2 shows the overview of the proposed network architecture, which can be trained end-to-end. Convolutional features of the image are sent to the iterative 3D regression module whose objective is to infer the 3D human body and the camera such that its 3D joints *project* onto the annotated 2D joints. The inferred parameters are also sent to an adversarial discriminator network whose task is to determine if the 3D parameters are real meshes from the *unpaired* data. This encourages the network to output 3D human bodies that lie on the manifold of human bodies and acts as a weak-supervision for *in-the-wild* images without ground truth 3D annotations. Due to the rich representation of the 3D mesh model, this data-driven prior can capture joint angle limits, anthropometric constraints (*e.g*. height, weight, bone ratios), and subsumes the geometric priors used by models that only predict 3D joint locations [34, 40, 51]. When ground truth 3D information is available, we may use it as an intermediate loss. In all, our overall objective is

$$L = \lambda(L_{\text{reproj}} + \mathbb{1}L_{\text{3D}}) + L_{\text{adv}} \tag{1}$$

where λ controls the relative importance of each objective, $\mathbb{1}$ is an indicator function that is 1 if ground truth 3D is available for an image and 0 otherwise. We show results with and *without* the 3D loss. We discuss each component in the following.

3.1. 3D Body Representation

We encode the 3D mesh of a human body using the Skinned Multi-Person Linear (SMPL) model [24]. SMPL is a generative model that factors human bodies into *shape* – how individuals vary in height, weight, body proportions – and *pose* – how the 3D surface deforms with articulation. The shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ is parameterized by the first 10 coefficients of a PCA shape space. The pose $\boldsymbol{\theta} \in \mathbb{R}^{3K}$ is modeled by relative 3D rotation of $K = 23$ joints in axis-angle representation. SMPL is a differentiable function that outputs a triangulated mesh with $N = 6980$ vertices, $M(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{3\times N}$, which is obtained by shaping the template body vertices conditioned on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, then articulating the bones according to the joint rotations $\boldsymbol{\theta}$ via forward kinematics, and finally deforming the surface with linear blend skinning. The 3D keypoints used for reprojection error, $X(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{3\times P}$, are obtained by linear regression from the final mesh vertices.

We employ the weak-perspective camera model and solve for the global rotation $R \in \mathbb{R}^{3\times 3}$ in axis-angle representation, translation $t \in \mathbb{R}^2$ and scale $s \in \mathbb{R}$. Thus the set of parameters that represent the 3D reconstruction

of a human body is expressed as a 85 dimensional vector $\Theta = \{\boldsymbol{\theta}, \boldsymbol{\beta}, R, t, s\}$. Given Θ, the projection of $X(\boldsymbol{\theta}, \boldsymbol{\beta})$ is

$$\hat{\mathbf{x}} = s\Pi(RX(\boldsymbol{\theta}, \boldsymbol{\beta})) + t, \tag{2}$$

where Π is an orthographic projection.

3.2. Iterative 3D Regression with Feedback

The goal of the 3D regression module is to output Θ given an image encoding ϕ such that the joint reprojection error

$$L_{\text{reproj}} = \Sigma_i ||v_i(\mathbf{x}_i - \hat{\mathbf{x}}_i)||_1, \tag{3}$$

is minimized. Here $\mathbf{x}_i \in \mathbb{R}^{2\times K}$ is the ith ground truth 2D joints and $v_i \in \{0, 1\}^K$ is the visibility (1 if visible, 0 otherwise) for each of the K joints.

However, directly regressing Θ in one go is a challenging task, particularly because Θ includes rotation parameters. In this work, we take inspiration from previous works [7, 9, 31] and regress Θ in an iterative error feedback (IEF) loop, where progressive changes are made recurrently to the current estimate. Specifically, the 3D regression module takes the image features ϕ and the current parameters Θ_t as an input and outputs the residual $\Delta\Theta_t$. The parameter is updated by adding this residual to the current estimate $\Theta_{t+1} = \Theta_t + \Delta\Theta_t$. The initial estimate Θ_0 is set as the mean $\bar{\Theta}$. In [7, 31] the estimates are rendered to an image space to concatenate with the image input. In this work, we keep everything in the latent space and simply concatenate the features $[\phi, \Theta]$ as the input to the regressor. We find that this works well and is suitable when differentiable rendering of the parameters is non-trivial.

Additional direct 3D supervision may be employed when paired ground truth 3D data is available. The most common form of 3D annotation is the 3D joints. Supervision in terms of SMPL parameters $[\boldsymbol{\beta}, \boldsymbol{\theta}]$ may be obtained through MoSh [23, 48] when raw 3D MoCap marker data is available. Below are the definitions of the 3D losses. We show results with and without using any direct supervision L_{3D}.

$$L_{\text{3D}} = L_{\text{3D joints}} + L_{\text{3D smpl}} \tag{4}$$

$$L_{\text{joints}} = ||(\mathbf{X_i} - \hat{\mathbf{X}}_\mathbf{i})||_2^2 \tag{5}$$

$$L_{\text{smpl}} = ||[\boldsymbol{\beta}_i, \boldsymbol{\theta}_i] - [\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\theta}}_i]||_2^2. \tag{6}$$

Both [7, 31] use a "bounded" correction target to supervise the regression output at each iteration. However this assumes that the ground truth estimate is always known, which is not the case in our setup where many images do not have ground truth 3D annotations. As noted by these approaches, supervising each iteration with the final objective forces the regressor to overshoot and get stuck in local minima. Thus we only apply L_{reproj} and L_{3D} on the final estimate Θ_T, but apply the adversarial loss on the estimate at every iteration Θ_t forcing the network to take corrective steps that are on the manifold of 3D human bodies.

3.3. Factorized Adversarial Prior

The reprojection loss encourages the network to produce a 3D body that explains the 2D joint locations, however anthropometrically implausible 3D bodies or bodies with gross self-intersections may still minimize the reprojection loss. To regularize this, we use a discriminator network D that is trained to tell whether SMPL parameters correspond to a real body or not. We refer to this as an adversarial prior as in [47] since the discriminator acts as a data-driven prior that guides the 3D inference.

A further benefit of employing a rich, explicit 3D representation like SMPL is that we precisely know the meaning of the latent space. In particular SMPL has a factorized form that we can take advantage of to make the adversary more data efficient and stable to train. More concretely, we mirror the shape and pose decomposition of SMPL and train a discriminator for shape and pose independently. The pose is based on a kinematic tree, so we further decompose the pose discriminators and train one for each joint rotation. This amounts to learning the angle limits for each joint. In order to capture the joint distribution of the entire kinematic tree, we also learn a discriminator that takes in all the rotations. Since the input to each discriminator is very low dimensional (10-D for $\boldsymbol{\beta}$, 9-D for each joint and $9K$-D for all joints), they can each be small networks, making them rather stable to train. All pose discriminators share a common feature space of rotation matrices and only the final classifiers are learned separately.

Unlike previous approaches that make *a priori* assumptions about the joint limits [5, 52], we do not predefine the degrees of freedom of the kinematic skeleton model. Instead this is learned in a data-driven manner through this factorized adversarial prior. Without the factorization, the network does not learn to properly regularize the pose and shape, producing visually displeasing results. The importance of the adversarial prior is paramount when no paired 3D supervision is available. Without the adversarial prior the network produces totally unconstrained human bodies as we show in section 4.3.

While mode collapse is a common issue in GANs [10] we do not really suffer from this because the network not only has to fool the discriminator but also has to minimize the reprojection error. The images contain all the modes and the network is forced to match them all. The factorization may further help to avoid mode collapse since it allows generalization to unseen body shape and poses combinations.

In all we train $K + 2$ discriminators. Each discriminator D_i outputs values between [0, 1], representing the probability that Θ came from the data. In practice we use the least square formulation [25] for its stability. Let E represent the encoder including the image encoder and the 3D module.

Then the adversarial loss function for the encoder is

$$\min L_{\text{adv}}(E) = \sum_i \mathbb{E}_{\Theta \sim p_E}[(D_i(E(I)) - 1)^2], \quad (7)$$

and the objective for each discriminator is

$$\min L(D_i) = \mathbb{E}_{\Theta \sim p_{\text{data}}}[(D_i(\Theta) - 1)^2] + \mathbb{E}_{\Theta \sim p_E}[D_i(E(I))^2]. \quad (8)$$

We optimize E and all D_is jointly.

3.4. Implementation Details

Datasets: The *in-the-wild* image datasets annotated with 2D keypoints that we use are LSP, LSP-extended [17] MPII [3] and MS COCO [22]. We filter images that are too small or have less than 6 visible keypoints and obtain training sets of sizes $1k$, $10k$, $20k$ and $80k$ images respectively. We use the standard train/test split of these datasets. All test results are obtained using the ground truth bounding box.

For the 3D datasets we use Human3.6M [16] and MPI-INF-3DHP [28]. We leave aside sequences from training Subject 8 of MPI-INF-3DHP as the validation set to tune hyper-parameters, and use the full training set for the final experiments. Both datasets are captured in a controlled environment and provide 150k training images with 3D joint annotations. For Human3.6M, we also obtain ground truth SMPL parameters for the training images using MoSh [23] from the raw 3D MoCap markers. The unpaired data used to train the adversarial prior comes from MoShing three MoCap datasets: CMU [6], Human3.6M training set [16] and the PosePrior dataset [2], which contains an extensive variety of extreme poses. These consist of 390k, 150k and 180k samples respectively.

All images are scaled to 224×224 preserving the aspect ratio such that the diagonal of the tight bounding box is roughly 150px (see [17]). The images are randomly scaled, translated, and flipped. Mini-batch size is 64. When paired 3D supervision is employed each mini-batch is balanced such that it consists of half 2D and half 3D samples. All experiments use all datasets with paired 3D loss unless otherwise specified.

The definition of the $K = 23$ joints in SMPL do not align perfectly with the common joint definitions used by these datasets. We follow [5, 20] and use a regressor to obtain the 14 joints of Human3.6M from the reconstructed mesh. In addition, we also incorporate the 5 face keypoints from the MS COCO dataset [22]. New keypoints can easily be incorporated with the mesh representation by specifying the corresponding vertex IDs[1]. In total the reprojection error is computed over $P = 19$ keypoints.

Architecture: We use the ResNet-50 network [15] for encoding the image, pretrained on the ImageNet classification task [39]. The ResNet output is average pooled, producing features $\phi \in \mathbb{R}^{2048}$. The 3D regression module consists of two fully-connected layers with 1024 neurons each with a dropout layer in between, followed by a final layer of 85D neurons. We use $T = 3$ iterations for all of our experiments. The discriminator for the shape is two fully-connected layers with 10, 5, and 1 neurons. For pose, $\boldsymbol{\theta}$ is first converted to K many 3×3 rotation matrices via the Rodrigues formula. Each rotation matrix is sent to a common embedding network of two fully-connected layers with 32 hidden neurons. Then the outputs are sent to $K = 23$ different discriminators that output 1-D values. The discriminator for overall pose distribution concatenates all $K * 32$ representations through another two fully-connected layers of 1024 neurons each and finally outputs a 1D value. All layers use ReLU activations except the final layer. The learning rates of the encoder and the discriminator network are set to 1×10^{-5} and 1×10^{-4} respectively. We use the Adam solver [18] and train for 55 epochs. Training on a single Titan 1080ti GPU takes around 5 days. The λs and other hyper-parameters are set through validation data on MPI-INF-3DHP dataset. Implementation is in Tensorflow [1].

4. Experimental Results

Although we recover much more than 3D skeletons, evaluating the result is difficult since no ground truth mesh 3D annotations exist for current datasets. Consequently we evaluate quantitatively on the standard 3D joint estimation task. We also evaluate an auxiliary task of body part segmentation. In Figure 1 we show qualitative results on challenging images from MS COCO [22] with occlusion, clutter, truncation, and complex poses. Note how our model recovers head and limb orientations. In Figure 3 we show results on the test set of Human3.6M, MPI-INF-3DHP, LSP and MS COCO at various error percentiles. Our approach recovers reasonable reconstructions even at 95th percentile error. Please see the project website[2] for more results. In all figures, results on the model trained with and without paired 2D-to-3D supervision are rendered in light blue and light pink colors respectively.

4.1. 3D Joint Location Estimation

We evaluate 3D joint error on Human3.6M, a standard 3D pose benchmark captured in a lab environment. We also compare with the more recent MPI-INF-3DHP [27], a dataset covering more poses and actor appearances than Human3.6M. While the dataset is more diverse, it is still far from the complexity and richness of in-the-wild images.

We report using several error metrics that are used for evaluating 3D joint error. Most common evaluations report the mean per joint position error (*MPJPE*) and *Reconstruc-*

[1]The vertex ids in 0-indexing are nose: 333, left eye: 2801, right eye: 6261, left ear: 584, right ear: 4072.

[2]https://akanazawa.github.io/hmr/

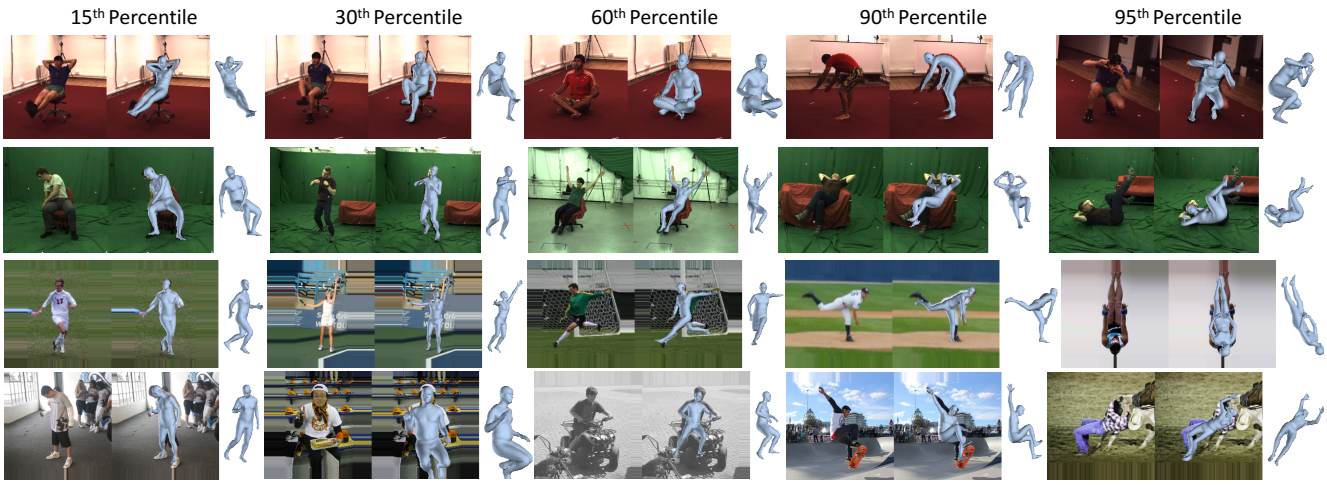

Figure 3: **Results sampled from different datasets at the 15th, 30th, 60th, 90th and 95th error percentiles.** Percentiles are computed using MPJPE for 3D datasets (first two rows - Human3.6M and MPI-INF-3DHP) and 2D pose PCK for 2D datasets (last two rows - LSP and MS COCO). High percentile indicates high error. Note results at high error percentile are often semantically quite reasonable.

tion error, which is MPJPE after rigid alignment of the prediction with ground truth via Procrustes Analysis [11]). Reconstruction error removes global misalignments and evaluates the quality of the reconstructed 3D skeleton.

Human3.6M We evaluate on two common protocols. The first, denoted P1, is trained on 5 subjects (S1, S5, S6, S7, S8) and tested on 2 (S9, S11). Following previous work [33, 36], we downsample all videos from 50fps to 10fps to reduce redundancy. The second protocol [5, 44], P2, uses the same train/test set, but is tested only on the frontal camera (camera 3) and reports reconstruction error.

We compare results for our method (HMR) on P2 in Table 1 with two recent approaches [5, 20] that also output SMPL parameters from a single image. Both approaches require 2D keypoint detection as input and we out-perform both by a large margin. We show results on P1 in Table 2. Here we also out-perform the recent approach of Zhou *et al.* [52], which also outputs 3D joint angles in a kinematic tree instead of joint positions. Note that they specify the DoF of each joint by hand, while we learn this from data. They also assume a fixed bone length while we solve for shape. HMR is competitive with recent state-of-the-art methods that only predict the 3D joint locations.

We note that MPJPE does not appear to correlate well with the visual quality of the results. We find that many results with high MPJPE appear quite reasonable as shown in Figure 3, which shows results at various error percentiles.

MPI-INF-3DHP The test set of MPI-INF-3DHP consists of 2929 valid frames from 6 subjects performing 7 actions. This dataset is collected indoors and outdoors with a multi-camera marker-less MoCap system. Because of this, the ground truth 3D annotations have some noise. In addition to MPJPE, we report the Percentage of Correct Keypoints (PCK) thresholded at 150mm and the Area Under the Curve (AUC) over a range of PCK thresholds [27].

Method	Reconst. Error
Rogez *et al.* [35]	87.3
Pavlakos *et al.* [33]	51.9
Martinez *et al.* [26]	**47.7**
*Regression Forest from 91 kps [20]	93.9
*SMPLify [5]	82.3
*SMPLify from 91 kps [20]	80.7
*HMR	**56.8**
*HMR unpaired	66.5

Table 1: **Human3.6M, Protocol 2.** Showing reconstruction loss (mm); * indicates methods that output more than 3D joints. HMR, with and *without* direct 3D supervision, out-performs previous approaches that output SMPL from 2D keypoints.

Method	MPJPE	Reconst. Error
Tome *et al.* [44]	88.39	
Rogez *et al.* [36]	87.7	71.6
VNect *et al.* [28]	80.5	
Pavlakos *et al.* [33]	71.9	**51.23**
Mehta *et al.* [27]	68.6	
Sun *et al.* [40]	**59.1**	
*Deep Kinematic Pose [52]	107.26	
*HMR	**87.97**	58.1
*HMR unpaired	106.84	67.45

Table 2: **Human3.6M, Protocol 1.** MPJPE and reconstruction loss in mm. * indicates methods that output more than 3D joints.

	Absolute			After Rigid Alignment		
Method	PCK	AUC	MPJPE	PCK	AUC	MPJPE
Mehta *et al.* [27]	75.7	39.3	**117.6**	-	-	-
VNect [28]	**76.6**	**40.4**	124.7	83.9	47.3	98.0
*HMR	72.9	36.5	124.2	**86.3**	**47.8**	**89.8**
*HMR unpaired	59.6	27.9	169.5	77.1	40.7	113.2

Table 3: **Results on MPI-INF-3DHP with and without rigid alignment.** * are methods that output more than 3D joints. Accuracy increases with alignment (PCK and AUC increase, while MPJPE decreases).

Method	Fg vs Bg		Parts		Run Time
	Acc	F1	Acc	F1	
SMPLify *oracle*[20]	92.17	0.88	88.82	0.67	-
SMPLify [5]	91.89	0.88	87.71	0.64	∼1 min
Decision Forests[20]	86.60	0.80	82.32	0.51	0.13 sec
HMR	91.67	0.87	87.12	0.60	**0.04 sec**
HMR unpaired	91.30	0.86	87.00	0.59	**0.04 sec**

Table 4: **Foreground and part segmentation (6 parts + bg) on LSP [20].** Reporting average accuracy and F1-score (higher the better). Proposed HMR is comparable to the oracle SMPLify which uses ground truth segmentation in fitting SMPL.

The results are shown in Table 3. All methods use the perspective correction of [27]. We also report metrics after rigid alignment for HMR and VNect using the publicly available code [28]. We report VNect results without post-processing optimization over time. Again, we are competitive with approaches that are trained to output 3D joints and we improve upon VNect after rigid alignment.

4.2. Human Body Segmentation

We also evaluate our approach on the auxiliary task of human body segmentation on the 1000 test images of LSP [17] labeled by [20]. The images have labels for six body part segments and the background. Note that LSP contains complex poses of people playing sports and no ground truth 3D labels are available for training. We do not use the segmentation label during training either.

We report the segmentation accuracy and average F1 score over all parts including the background as done in [20]. We also report results on foreground-background segmentation. Note that the part definition segmentation of the SMPL mesh is not exactly the same as that of annotation; this limits the best possible accuracy to be less than 100%.

Results are shown in Table 4. Our results are comparable to the SMPLify oracle [20], which uses ground truth segmentation and keypoints as the optimization target. It also out-performs the Decision Forests of [20]. Note that HMR is also real-time given a bounding box.

Figure 4: **Results with and without paired 3D supervision.** 3D reconstructions, without direct 3D supervision, are very close to those of the supervised model.

Figure 5: **No Discriminator No 3D.** With neither the discriminator, nor the direct 3D supervision, the network produces monsters. On the right of each example we visualize the ground truth keypoint annotation in unfilled circles, and the projection in filled circles. Note that despite the unnatural pose and shape, its 2D projection error is very accurate.

4.3. Without Paired 3D Supervision

So far we have used paired 2D-to-3D supervision, *i.e.* L_{3D} whenever available. Here we evaluate a model trained without any paired 3D supervision. We refer to this setting as *HMR unpaired* and report numerical results in all the tables. All methods that report results on the 3D joint estimation task rely on direct 3D supervision and cannot train without it. Even methods that are based on a reprojection loss [33, 47, 50] require paired 2D-to-3D training data.

The results are surprisingly competitive given this challenging setting. Note that the adversarial prior is essential for training without paired 2D-to-3D data. Figure 5 shows that a model trained with neither the paired 3D supervision nor the adversarial loss produces monsters with extreme shape and poses. It remains open whether increasing the amount of 2D data will significantly increase 3D accuracy.

5. Conclusion

In this paper we present an end-to-end framework for recovering a full 3D mesh model of a human body from a single RGB image. We parameterize the mesh in terms of 3D joint angles and a low dimensional linear shape space, which has a variety of practical applications. In this past few years there has been rapid progress in single-view 3D pose prediction on images captured in a controlled environment. Although the performance on these benchmarks is starting to saturate, there has not been much progress on 3D human reconstruction from images *in-the-wild*. Our results without using any paired 3D data are promising since

they suggest that we can keep on improving our model using more images with 2D labels, which are relatively easy to acquire, instead of ground truth 3D, which is considerably more challenging to acquire in a natural setting.

Acknowledgements. We thank N. Mahmood for the SMPL model fits to mocap data and the mesh retargeting for character animation, D. Mehta for his assistance on MPI-INF-3DHP, and S. Tulsiani, A. Kar, S. Gupta, D. Fouhey and Z. Liu for helpful discussions. This research was supported by BAIR sponsors and NSF Award IIS-1526234.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *Operating Systems Design and Implementation*, 2016. 6

[2] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, June 2015. 3, 6

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, June 2014. 6

[4] C. Barron and I. Kakadiaris. Estimating anthropometry and pose from a single uncalibrated image. *Computer Vision and Image Understanding, CVIU*, 81(3):269–284, 2001. 3

[5] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision, ECCV*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 2, 3, 4, 5, 6, 7, 8

[6] F. by NSF EIA-0196217. Cmu graphics lab - motion capture library. http://mocap.cs.cmu.edu/. 6

[7] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016. 5

[8] Y. Chen, T.-K. Kim, and R. Cipolla. Inferring 3D shapes and deformations from single views. In *European Conference on Computer Vision, ECCV*, pages 300–313, 2010. 3

[9] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1078–1085. IEEE, 2010. 5

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 5

[11] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, Mar 1975. 7

[12] P. Guan, A. Weiss, A. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *IEEE International Conference on Computer Vision, ICCV*, pages 1381–1388, 2009. 3

[13] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017. 4

[14] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormhlen, and H. P. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1823–1830, 2010. 3

[15] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision, ECCV*, 2016. 6

[16] 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*, 36(7):1325–1339, 2014. 3, 6

[17] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, pages 12.1–12.11, 2010. 6, 8

[18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, and V. Mansinghka. Picture: A probabilistic programming language for scene perception. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4390–4399, 2015. 4

[20] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, July 2017. 2, 3, 4, 6, 7, 8

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

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, Zrich, 2014. Oral. 4, 6

[23] M. Loper, N. Mahmood, and M. J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia*, 33(6):220:1–220:13, 2014. 5, 6

[24] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 4

[25] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks, 2016. 5

[26] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision, ICCV*, 2017. 3, 7

[27] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *Proc. of International Conference on 3D Vision (3DV)*, 2017. 3, 6, 7, 8

[28] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH*, 36, July 2017. 3, 4, 6, 7, 8

[29] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. *arXiv preprint arXiv:1611.09010*, 2016. 3

[30] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision, ECCV*, pages 483–499, 2016. 3

[31] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3316–3324, 2015. 5

[32] V. Parameswaran and R. Chellappa. View independent human body pose estimation from a single perspective image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 16–22, 2004. 3

[33] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017. 3, 7, 8

[34] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d Human Pose from 2d Image Landmarks. *Computer Vision–ECCV 2012*, pages 573–586, 2012. 3, 4

[35] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Advances in Neural Information Processing Systems*, pages 3108–3116, 2016. 3, 7

[36] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net: Localization-Classification-Regression for Human Pose. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, July 2017. 3, 7

[37] M. Sanzari, V. Ntouskos, and F. Pirri. Bayesian image based 3d pose estimation. In *European Conference on Computer Vision, ECCV*, pages 566–582, 2016. 3

[38] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision, IJCV*, 87(1):4–27, 2010. 3

[39] N. Silberman and S. Guadarrama. Tensorflow-slim image classification model library. https://github.com/tensorflow/models/tree/master/research/slim. 6

[40] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *IEEE International Conference on Computer Vision, ICCV*, 2017. 3, 4, 7

[41] J. K. V. Tan, I. Budvytis, and R. Cipolla. Indirect deep structured learning for 3d human shape and pose prediction. In *Proceedings of the British Machine Vision Conference*, 2017. 4

[42] C. Taylor. Reconstruction of articulated objects from point correspondences in single uncalibrated image. *Computer Vision and Image Understanding, CVIU*, 80(10):349–363, 2000. 2, 3

[43] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua. Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation. In *IEEE International Conference on Computer Vision, ICCV*, 2017. 3

[44] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3, 7

[45] S. Tulsiani and J. Malik. Viewpoints and keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015. 3

[46] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5242–5252, 2017. 4

[47] H.-Y. F. Tung, A. W. Harley, W. Seto, and K. Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *IEEE International Conference on Computer Vision, ICCV*, 2017. 3, 5, 8

[48] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from Synthetic Humans. In *CVPR*, 2017. 4, 5

[49] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4724–4732, 2016. 3

[50] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *European Conference on Computer Vision, ECCV*, 2016. 3, 8

[51] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Weakly-supervised transfer for 3d human pose estimation in the wild. In *IEEE International Conference on Computer Vision, ICCV*, 2017. 3, 4

[52] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCV Workshop on Geometry Meets Deep Learning*, pages 186–201, 2016. 3, 4, 5, 7

[53] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparse representation for 3D shape estimation: A convex relaxation approach. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4447–4455, 2015. 3

[54] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4966–4975, 2016. 3

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017. 4