# Learning to Estimate 3D Human Pose and Shape from a Single Color Image

Georgios Pavlakos[1], Luyang Zhu[2], Xiaowei Zhou[3], Kostas Daniilidis[1]
[1] University of Pennsylvania [2] Peking University [3] State Key Lab of CAD&CG, Zhejiang University

## Abstract

*This work addresses the problem of estimating the full body 3D human pose and shape from a single color image. This is a task where iterative optimization-based solutions have typically prevailed, while Convolutional Networks (ConvNets) have suffered because of the lack of training data and their low resolution 3D predictions. Our work aims to bridge this gap and proposes an efficient and effective direct prediction method based on ConvNets. Central part to our approach is the incorporation of a parametric statistical body shape model (SMPL) within our end-to-end framework. This allows us to get very detailed 3D mesh results, while requiring estimation only of a small number of parameters, making it friendly for direct network prediction. Interestingly, we demonstrate that these parameters can be predicted reliably only from 2D keypoints and masks. These are typical outputs of generic 2D human analysis ConvNets, allowing us to relax the massive requirement that images with 3D shape ground truth are available for training. Simultaneously, by maintaining differentiability, at training time we generate the 3D mesh from the estimated parameters and optimize explicitly for the surface using a 3D per-vertex loss. Finally, a differentiable renderer is employed to project the 3D mesh to the image, which enables further refinement of the network, by optimizing for the consistency of the projection with 2D annotations (i.e., 2D keypoints or masks). The proposed approach outperforms previous baselines on this task and offers an attractive solution for direct prediction of 3D shape from a single color image.*

## 1. Introduction

Estimating the full body 3D pose and shape of humans from images has been a challenging goal of computer vision going all the way back to the work of Hogg [15]. The inherent ambiguity of the problem has forced the researchers to use monocular image sequences for inference [54, 3], employ multiple camera views [36, 16], or even explore alternative sensors, like Kinect [53] or IMUs [52]. In these settings, the body shape reconstruction results are remarkable. However, estimating 3D pose and shape from single color images remains the ultimate goal for 3D human analysis.

Considering the particularly challenging nature of such a problem, the literature remains undeniably sparse. Most approaches rely on iterative optimization, attempting to estimate a full body 3D shape that is consistent with 2D image observations, like silhouettes, edges, shading, or 2D keypoints [41, 14]. Despite the significant runtime required to solve the complicated optimization problem, the common failures because of local minima, and the error-prone reliance on ambiguous 2D cues, optimization-based solutions remain the leading paradigm for this problem [22, 7]. Even the emergence of deep learning has not changed significantly the landscape. ConvNets did not seem as a viable candidate for this problem because they require a huge amount of training data and they are infamous for their low resolution 3D predictions [37, 44]. The goal of our work is to demonstrate that ConvNets can indeed offer an attractive solution for this problem, by proposing an efficient and effective direct prediction approach, which is competitive and even outperforms iterative optimization methods.

To make this feasible, a critical design choice for our approach is the incorporation of a parametric statistical body shape model (SMPL [25]) within our end-to-end framework, presented in Figure 1. The advantage of such a representation is that we can generate high quality 3D meshes in the form of 6890 vertices while estimating only a small number of parameters, i.e., 72 for pose and 10 for shape. This low-dimensional parameterization makes the model friendly for direct network prediction. In fact, this prediction is feasible and accurate by using only 2D keypoints and silhouettes as input. This allows us to relax the limiting assumption that natural images with 3D shape ground truth are available for training. In contrast, we can leverage the available 2D image annotations (e.g., [19, 4]) to train for image-to-2D inference, while using instances of the parametric model to train for 2D-to-3D shape inference. Simultaneously, another major advantage of employing this parametric model is that its structure allows us to generate the estimated 3D mesh at training time and optimize directly for the surface, by using a 3D per-vertex loss. This loss has better correlation with the vertex-to-vertex 3D error that is typically used for evaluation and improves training compared to

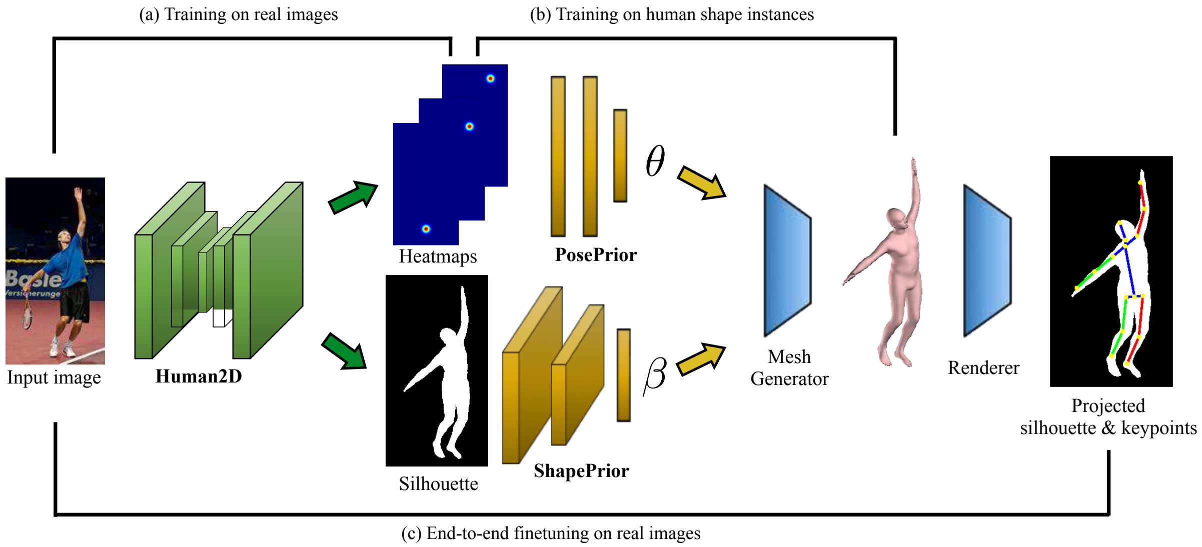

Figure 1. Schematic representation of our framework. (a) An initial ConvNet, *Human2D*, predicts 2D heatmaps and masks from a single color image, using 2D pose data [19, 4] for training. (b) Two networks estimate the parameters of the statistical model SMPL [25], using instances of the parametric model for training. The *PosePrior* estimates pose parameters ($\theta$) from keypoints, and the *ShapePrior* estimates shape parameters ($\beta$) from silhouettes. (c) The framework can be finetuned end-to-end without requiring images with 3D shape ground truth, by projecting the full body 3D mesh to the image and optimizing for the consistency of the projection with 2D annotations (keypoints and masks). The blue parts (Mesh Generator and Renderer) indicate components without learnable parameters.

naive parameter regression. Finally, we propose to employ a differentiable renderer to project the generated 3D mesh back to the 2D image. This enables end-to-end finetuning of the network by optimizing for the consistency of the projection with annotated 2D observations, i.e., 2D keypoints and masks. The complete framework offers a modular direct prediction solution to the problem of 3D human pose and shape estimation from a single color image and outperforms previous approaches on the relevant benchmarks.

Our main contributions can be summarized as follows:

- an end-to-end framework for 3D human pose and shape estimation from a single color image.
- incorporation of a parametric statistical shape model, SMPL, within the end-to-end framework, enabling:
  - prediction of the SMPL model parameters from ConvNet-estimated 2D keypoints and masks to avoid training on synthetic image examples.
  - generation of the 3D body mesh at training time and supervision based on the 3D shape consistency.
  - use of a differentiable renderer for 3D mesh projection and refinement of the network with supervision based on the consistency with 2D annotations.
- superior performance compared to previous approaches for 3D human pose and shape estimation at significantly faster running time.

## 2. Related work

**3D human pose estimation**: In order to estimate a convincing 3D reconstruction of the human body, it is crucial to get an accurate prediction of the 3D pose of the person. Many recent works follow the end-to-end paradigm [48, 40, 42, 46, 55], using images as input to predict 3D joint locations [23, 45, 34, 28], regress 3D heatmaps [31], or classify the image in a particular pose class [39, 40]. Unfortunately, an important constraint is that most of these ConvNets require images with 3D pose ground truth for training, limiting the available training data sources. Other approaches commit to the 2D pose estimates provided by state-of-the-art ConvNets and focus on the 3D pose reconstruction [29, 57], recover 3D pose exemplars [8], or produce multiple 3D pose candidates consistent with the 2D pose [18]. Notably, Martinez *et al.* [27] demonstrate state-of-the-art results using a simple multi-layer perceptron which regresses the 3D joint locations from 2D pose input. Our goal is significantly different from the aforementioned works, since instead of a rough stickman-like figure, we estimate the whole surface geometry of the human body.

**Human shape estimation**: Concurrently with advances in 3D human pose, a different set of works addressed the problem of human shape estimation. In this case, given a single image, most methods attempt to estimate the parameters of a statistical body shape model like SCAPE [5] or SMPL [25]. The input is usually silhouettes, while regression forests [9] and ConvNets [11, 10] have been proposed for the prediction. Knowledge of human shape is useful for biometric applications, however we argue that for 3D perception the potential and the challenges are significantly greater when pose and shape are inferred jointly.

**Joint 3D human pose and shape estimation**: Despite individual advances in pose and shape prediction, their joint estimation makes the task significantly harder. This has consistently fostered research in non single image scenarios, for more robust results. Xu *et al.* [54] propose a pipeline for full performance capture from monocular video assuming knowledge of the shape mesh for the observed subject. Alldieck *et al.* [3] estimate pose and shape jointly from monocular video relying on optical flow cues. Rhodin *et al.* [36] and Huang *et al.* [16] use images from multiple calibrated cameras and rely on keypoint detections, silhouettes and temporal consistency to recover a reconstruction of the body. An alternative setting is proposed by Weiss *et al.* [53] making use of the depth modality of the Kinect sensor to tackle the same problem. In the same spirit of exploring different sensors, von Marcard *et al.* [52] use a sparse set of IMUs on the subject to recover pose and shape jointly.

**3D human pose and shape from a single color image**: In the most challenging case of using only a single color image as input, the work of Sigal *et al.* [41] is among the first to estimate high quality 3D shape estimates, by fitting the parametric model SCAPE [5] to ground truth image silhouettes. Guan *et al.* [14] use silhouettes, edges and shading as cues during the fitting process, but still require initialization through a user specified 2D skeleton. A fully automatic approach was proposed very recently by Bogo *et al.* [7]. They use 2D keypoint detections from a 2D pose ConvNet [33] and fit the parametric model SMPL [25] to these 2D locations. Their 3D pose results are very accurate, but shape remains highly underconstrained. To improve upon this, Lassner *et al.* [22] extends the fitting using silhouettes provided by a segmentation ConvNet. The common theme of these works is that they pose an optimization problem and attempt to fit a body model to a set of 2D observations. The drawback though is that solving this iterative optimization problem is very slow, it can easily fail because of local minima, and it relies a lot on error-prone 2D observations.

Alternatively, direct prediction approaches estimate 3D pose and shape in a discriminative way, without explicitly optimizing a specific objective during inference. Relevant to this paradigm is the work of Lassner *et al.* [22], where a ConvNet detects 91 landmarks of the human body and then a random forest estimates the 3D body and shape from these detections. However, to train for these landmarks, they still require alignment of body shapes with images. In contrast, we demonstrate that only a much smaller set of annotations are critical for the reconstruction, i.e., 2D joints and masks, which can be provided by human annotators and are abundant for in-the-wild images [19, 4, 24], while we also incorporate everything within a unified end-to-end framework. Concurrently, Tan *et al.* [43] use an encoder-decoder ConvNet, where the decoder is trained to predict the silhouette corresponding to SMPL parameters. We differ to them by identifying that from these parameters we can analytically generate the body mesh and project it to the image in a differentiable way (as in [47] for face models), avoiding half a million of extra learnable weights. Instead, we focus our computational and learning effort in the image to 3D shape part of the framework. Our work is also related to the concurrent work of Tung *et al.* [50], however our framework can be trained from scratch instead of relying on synthetic image data for pretraining, and we demonstrate state-of-the-art results for model-based 3D pose and shape prediction.

## 3. Human body shape models

Statistical body shape models, like SCAPE [5] or SMPL [25], are powerful tools, which provide significant opportunities for an end-to-end framework. One of the important advantages is their low-dimensional parameter space, which is very suitable for direct network prediction. With this parameter representation, we can keep the output prediction space small, compared to voxelized or point cloud representations. Simultaneously, the low dimensional prediction does not sacrifice the quality of the output, since we can still generate high quality 3D meshes from the estimated parameters. Furthermore, from a learning perspective, we bypass the problem of learning the statistics of the human body, and devote the network capacity at the inference of the model parameters from image evidence. In contrast, approaches without the aid of a model put additional burden on the learning side, which often leads to embarrassing prediction errors (e.g., failing to reconstruct limbs under occlusion, missing body details, etc). Moreover, most models offer a convenient disentanglement of pose and shape which is useful to independently focus on the factors that affect each one of the two. Last but certainly not least for end-to-end approaches, the function which generates the 3D mesh from parameter inputs is differentiable, making the models compatible with current end-to-end pipelines.

In this work, we employ the more recent SMPL model, introduced by Loper *et al.* [25]. We provide the essential notation here, and we refer the reader to [25] for more details. SMPL defines a function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}; \Phi)$, where $\boldsymbol{\beta}$ are the shape parameters, $\boldsymbol{\theta}$ are the pose parameters and $\Phi$ are fixed parameters of the model. The direct output of this function is a body mesh $\boldsymbol{P} \in \mathbb{R}^{N \times 3}$ with $N = 6890$ vertices $P_i \in \mathbb{R}^3$. The shape of the model uses a linear combination of a low number of principal body shapes which are learned from a large dataset of body scans [38]. The *shape parameters* $\boldsymbol{\beta}$ are the linear coefficients of these base shapes. The pose of the body is defined through a skeleton rig with 23 joints. The *pose parameters* $\boldsymbol{\theta}$ are expressed in the axis angle representation and define the relative rotation between parts of the skeleton. In total, 72 parameters define the pose (3 for each of the 23 joints, plus 3 for the global rotation). Given the rest pose shape retrieved by the shape parameters $\beta$, SMPL defines pose-dependent deformations and uses the pose parameters $\boldsymbol{\theta}$ to produce the final output mesh. Conveniently, the *body joints* $\boldsymbol{J}$ are a linear combination of a sparse set of mesh vertices, making joints a direct outcome of the estimated body mesh.

## 4. Technical approach

The conventional ConvNet-based approach for our task would be to acquire a large amount of color images with 3D shape ground truth and train the network with these input-output pairs. However, except for small-scale datasets [22] or synthetically generated image examples [51] this type of data is typically unavailable. Therefore, to deal with this task, we need to rethink the typical pipeline. Our main goal is to leverage all the resources we have available and use our insights for the problem to build an effective framework. As a first step, from findings of prior work, we identify that 3D pose can be estimated reliably from 2D pose estimates [7, 27], while the shape can be inferred from silhouette measurements [11, 10]. This observation conveniently decomposes the problem in a) estimation of keypoints and masks from color images and, b) prediction of 3D pose and shape from the 2D evidence. The advantage of this practice is that the framework can be trained without requiring images with 3D shape ground truth.

### 4.1. Keypoints and silhouette prediction

The first step of our framework focuses on 2D keypoint and silhouette estimation. This part is motivated by the availability of large-scale benchmarks [19, 4, 24] with 2D joints and mask annotations. Considering the volume and the variability of this data, we leverage it to train a ConvNet for 2D pose and silhouette prediction, that is particularly reliable under various imaging conditions and poses.

In the past, two individual ConvNets have been used to provide 2D keypoints and masks [16, 22]. In contrast, for a more elegant solution, we train a single ConvNet, which we denote as *Human2D*, that generates two outputs, one for keypoints and one for silhouettes. *Human2D* follows the Stacked Hourglass design [30], using two hourglasses, which was found to be a good trade-off between accuracy and running time. The keypoint output is in the form of heatmaps [49, 32], where an MSE loss, $\mathcal{L}_{hm}$, between the ground truth and the predicted heatmaps is used for supervision. The silhouette output has two channels (body and background) and is supervised using a pixelwise binary cross entropy loss, $\mathcal{L}_{sil}$. For training, we combine the two losses: $\mathcal{L}_{hg} = \lambda\mathcal{L}_{hm} + \mathcal{L}_{sil}$, where $\lambda = 100$. This ConvNet falls under the multi-task learning paradigm [34]. Through sharing, the two tasks might benefit each other, but multi-task learning can also pose certain challenges (e.g., appropriate weighting of the losses), as Kokkinos identifies [21].

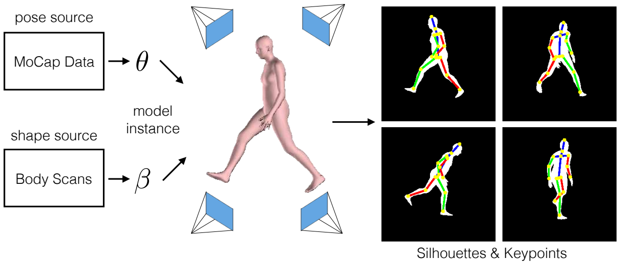

Figure 2. We aim to learn the mapping from silhouettes and keypoints to model parameters, so we can synthesize body model instances and project them to the image plane to simulate the network input. We only require a source to sample pose parameters, and a source to sample body shape parameters. Projections from different viewpoints can also be employed for data augmentation.

### 4.2. 3D pose and shape prediction

The second step is significantly more challenging, requiring estimation of the full body 3D pose and shape from 2D keypoints and silhouettes. Silhouettes and/or keypoints have been used extensively for 3D model fitting through iterative optimization [6, 7, 22]. Here, we demonstrate that this mapping can also be learned from data while it is possible to get a reliable prediction in a single estimation step.

For this mapping, we train two network components: (a) the *PosePrior*, which uses 2D keypoint locations as input together with the confidence of the detections (realised by the maximum value of each heatmap) and estimates the pose coefficients $\boldsymbol{\theta}$, and (b) the *ShapePrior*, which uses the silhouette as input and estimates the shape coefficients $\boldsymbol{\beta}$. In general, the silhouette can be helpful for 3D pose inference [6] and vice versa [7]. However, empirically we discovered this disentanglement to provide more stable and accurate 3D predictions, while it also leads to a more modular pipeline (e.g. updating only the *PosePrior*, without retraining the whole network). Regarding the architecture, the *PosePrior* uses two bilinear units [27], where the input is the 2D keypoint locations and the maximum responses from each heatmap, and the output is the 72 SMPL pose parameters $\boldsymbol{\theta}$. The ShapePrior uses a simple architecture with five $3 \times 3$ convolutional layers, each one followed by max-pooling, and an additional bilinear unit at the end with 10 outputs, corresponding to the SMPL shape parameters $\boldsymbol{\beta}$.

The form of the input (2D keypoints and masks) and the output (shape and pose parameters) allows us to produce large amount of training data by generating instances of the SMPL model with different 3D pose and shape (Figure 2). In fact, we can leverage MoCap data (e.g., [1, 17]) to sample 3D poses, and body scans (e.g., [38]) to sample body shapes. For the input, we only need to project the 3D model to the image plane (possibly from different viewpoints), and compute silhouettes and 2D keypoint locations to generate input-output pairs for training. This data generation is feasible, exactly because we used an intermediate silhouette and keypoints representation. In contrast, attempting to learn a mapping directly from color images would require generation of synthetic image examples [51], which typically do not reach the variability of in-the-wild images.

In the previous paragraphs, we deliberately avoided discussing the supervision of the *Priors* networks. Past works [22, 43] have examined supervision schemes using a typical $\mathcal{L}_2$ loss between the predicted and ground truth parameters. One shortcoming of this naive parameter regression approach, is that different parameters might have effects of different scale on the final reconstruction (e.g., the global body rotation is much more crucial than the local rotation of the hand with respect to the wrist). To avoid hand-selecting or tuning the supervision for each parameter, we aim for a more global solution. Our approach entails the generation of the full body mesh at training time, where we optimize explicitly for the predicted surface by applying a 3D per-vertex loss. Since the function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}; \Phi)$ is differentiable, we can backpropagate through it and handle this mesh generator as a typical layer of our network, without any learnable parameters. Given the predicted mesh vertices $\hat{P}_i$ and the corresponding groundturth vertices $P_i$, we can supervise the network with a *3D per-vertex loss*:

$$\mathcal{L}_{\mathcal{M}} = \sum_{i=1}^{N} \|\hat{P}_i - P_i\|_2^2, \tag{1}$$

which considers all the vertices equally and has better correlation with the 3D per-vertex error which is usually employed for evaluation. Alternatively, if the focus is mainly on 3D pose, we can also supervise the network considering only the $M$ relevant 3D joints $J_i$, which are trivially exposed by the model as a sparse linear combination of the mesh vertices. In this case, denoting with $\hat{J}_i$ the estimated joints, the corresponding loss can be expressed as:

$$\mathcal{L}_{\mathcal{J}} = \sum_{i=1}^{M} \|\hat{J}_i - J_i\|_2^2. \tag{2}$$

Empirically, we found that the best training strategy is to initially get a reasonable initialization for the network parameters using an $\mathcal{L}_2$ parameter loss, and then activate also the vertex loss $\mathcal{L}_{\mathcal{M}}$ (or the joints loss $\mathcal{L}_{\mathcal{J}}$ if the focus is on pose only), to train a better model.

### 4.3. Differentiable renderer

Our previous analysis relaxed the assumption that images with 3D shape ground truth are available for training and relied on geometric 3D data (MoCap and body scans). In some cases though, even this type of data might be unavailable. For example, LSP [19] has gymnastics or parkour poses which are not represented in typical MoCap. Luckily, our generated 3D mesh has potential to leverage these 2D annotations for training purposes.

To close the loop, our complete approach includes an additional step that projects the 3D mesh to the image and examines consistency with 2D annotations. In concurrent work, a decoder-type network was used to learn the mapping from SMPL parameters to silhouettes [43]. However, here we identify that this mapping is known and involves the projection of the 3D mesh to the image, which can be expressed in a differentiable way, without the need to train a network with learnable weights. More specifically, for our implementation, we employ an approximately differentiable renderer, OpenDR [26], which projects the mesh and the 3D joints to the image space, and enables backpropagation. The projection operation $\Pi$ gives rise to: (a) the silhouette $\Pi(\hat{\boldsymbol{P}}) = \hat{S}$, which is represented as a $64 \times 64$ binary image, and (b) the projected 2D joints $\Pi(\hat{\boldsymbol{J}}) = \hat{\boldsymbol{W}} \in \mathbb{R}^{M \times 2}$. In this case, the supervision comes from the comparison of these projections with the annotated silhouettes $S$, and the 2D keypoints $\boldsymbol{W}$, using $\mathcal{L}_2$ losses:

$$\mathcal{L}_{\Pi} = \mu \sum_{i}^{M} \|\hat{W}_i - W_i\|_2^2 + \|\hat{S} - S\|_2^2, \tag{3}$$

where $\mu = 10$. The goal of this type of supervision is twofold: (a) it can be employed for end-to-end refinement of the network, using only images with 2D keypoints and/or masks for training, and (b) it can be useful to mildly adapt a generic pose or shape prior to a new setting (e.g., new dataset), where only 2D annotations are available.

## 5. Empirical evaluation

This section focuses on the empirical evaluation of the proposed approach. First, we present the benchmarks that we employed for quantitative and qualitative evaluation. Then, we provide some essential implementation details of the approach. Finally, quantitative and qualitative results are presented on the selected datasets.

### 5.1. Datasets

For the empirical evaluation, we employed two recent benchmarks that provide color images with 3D body shape ground truth, the UP-3D dataset [22] and the SURREAL dataset [51]. Additionally, we used the Human3.6M [17] dataset for further evaluation of the 3D pose accuracy.

**UP-3D**: It is a recent dataset that collects color images from 2D human pose benchmarks, like LSP [19] and MPII [4] and uses an extended version of SMPLify [7] to provide 3D human shape candidates. The candidates were evaluated by human annotators to select only the images with good

3D shape fits. It comprises 8515 images, where 7818 are used for training and 1389 for testing. We report results on this test set, while we also consider subsets, based on the original dataset (LSP, MPII, or FashionPose) of the UP-3D images. Finally, we examine a reduced test set of 139 images, selected by Tan *et al.* [43] aiming to limit the range for the global rotation. We report results using the mean per-vertex error, between predicted and ground truth shape.

**SURREAL**: It is a recent dataset which provides synthetic image examples with 3D shape ground truth. The dataset draws poses from MoCap [1, 17] and body shapes from body scans [38] to generate valid SMPL instances for each image. The synthetic images are not very realistic, but the accurate ground truth, makes it a useful benchmark for evaluation. We report results on the Human3.6M part of the dataset, considering all test videos and keeping every fifth frame of each video to avoid excessive redundancy in the data. Results are reported using the mean per-vertex error.

**Human3.6M**: It is a large-scale indoor dataset that contains multiple subjects performing typical actions like “Eating” and “Walking”. We follow the protocol of Bogo *et al.* [7] using all videos of subjects S9 and S11 from ‘cam3’ for evaluation. The original videos are downsampled from 50fps to 10fps to remove redundancy as is done in [22]. Results are reported using the reconstruction error.

## 5.2. Implementation details

The *Human2D* network is trained on MPII [4], LSP [19] and LSP-extended [20] data, using the silhouettes from Lassner *et al.* [22]. We use a batch size of 4, learning rate set to 3e-4, and rmsprop for the optimization. Augmentation for rotation (±30°), scale (0.75-1.25) and flipping (left-right) is used. The training lasts for 1.2M iterations.

For the *Priors* networks, we train with a batch size of 256, learning rate set to 3e-4, and using rmsprop for the optimization. Initially, the networks are trained for 40k iterations using an $\mathcal{L}_2$ parameter loss, and then for 60k more iterations using also $\mathcal{L}_\mathcal{M}$ (or $\mathcal{L}_\mathcal{J}$ if we focus on pose only) weighted equally with the parameter loss.

The end-to-end refinement with the reprojection loss lasts for 2k iterations with a batch size of 4, learning rate set to 8e-5, and using rmsprop for the optimization. To improve training robustness, the end-to-end updates are alternated with individual updates of the *Human2D* and the *Priors* networks (as described in the previous two paragraphs). This helps the individual components to maintain their original purpose, while we are also leveraging the strength of end-to-end training to integrate them together.

## 5.3. Component evaluation

In this section, we evaluate the components of our approach, using the UP-3D dataset. We train two different versions of our system, where for *Priors* we leverage data either from UP-3D (provided by Lassner *et al.* [22]), or from CMU MoCap (provided by Varol *et al.* [51]). The *Human2D* network remains the same in both cases.

|  | Avg error |  |
|---|---|---|
| Data source for *Priors* | UP-3D | CMU |
| Parameter loss (axis-angle) | 514.9 | 589.9 |
| Parameter loss (rot matrix) | 140.7 | 152.2 |
| + Per-vertex loss | 120.7 | 142.0 |
| + Reprojection finetuning | 117.7 | 135.5 |

Table 1. Ablative study on UP-3D, comparing the different supervision forms on the same architecture. The numbers are mean per-vertex errors (mm). Two versions of the *Priors* networks are used, trained with data from UP-3D [22] and CMU [51] respectively. All networks are trained for the same number of iterations.

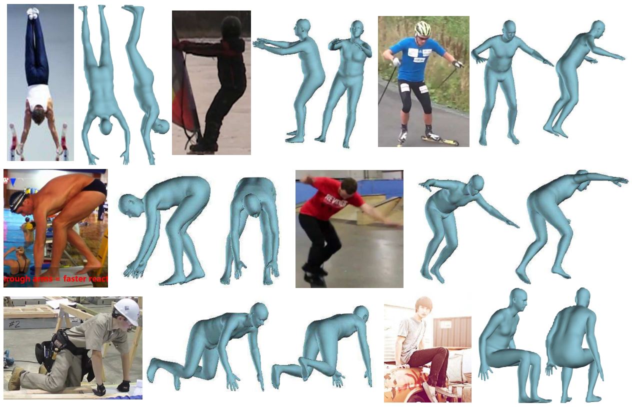

Figure 3. Successful 3D pose and shape predictions of our approach on challenging examples of UP-3D.

Our experiment focuses on the type of supervision. Naively training the *Priors* networks using an $\mathcal{L}_2$ loss for the $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ parameters [43], keeps the prediction error high as can be seen in Table 1 (line 1). Alternatively, we can transform the $\boldsymbol{\theta}$ parameters from axis-angle representation to rotation matrix using the Rodrigues' rotation formula [12], and apply an $\mathcal{L}_2$ loss on this representation instead (line 2). This leads to more stable training and better performance, as has also been observed by Lassner *et al.* [22]. However, generating the body mesh and further training of the network using our proposed per-vertex supervision (line 3) is even more appropriate and elevates our framework to state-of-the-art performance (see Section 5.4). Finally, the additional end-to-end finetuning with 2D annotations and the reprojection error (line 4) offers a mild refinement to the network. In the UP-3D case, the benefit is small, since the *Priors* have already observed very similar examples with full 3D ground truth, so 2D annotations become redundant. However, when training the *Priors* with CMU data, the domain shift, from CMU poses to UP-3D poses is significant, so these 2D annotations offers a clear

|  | LSP | MPII | Fashion | Full | Reduced |
|---|---|---|---|---|---|
| Lassner *et al.* [22] | 174.4 | 184.3 | 108.0 | 169.8 | 123.6 |
| Tan *et al.* [43] (Indirect) | - | - | - | - | 189 |
| Tan *et al.* [43] (Direct) | - | - | - | - | 105 |
| Ours | **127.8** | **110.0** | **106.5** | **117.7** | **100.5** |

Table 2. Detailed results on UP-3D [22]. The numbers are mean per vertex errors (mm), except for the 'Reduced' column where only 91 landmarks [22] contribute to the error. Our approach outperforms the other baselines across the table.

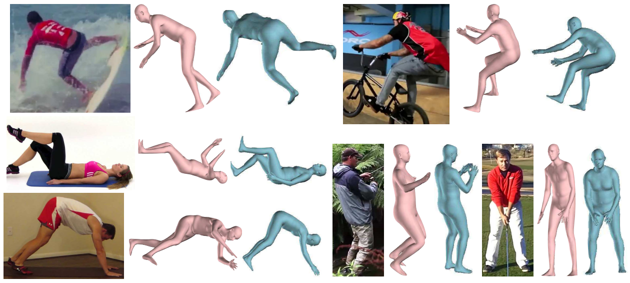

Figure 4. Examples from UP-3D where our approach (blue shapes) performs significantly better than the direct prediction method of Lassner *et al.* [22] (pink shapes).

performance benefit. This is an interesting empirical result demonstrating that training with reprojection losses can be useful not only for end-to-end refinement, but it can also assist the network with novel information recovered from 2D annotations. Some qualitative results from UP-3D using our best model are presented in Figure 3.

### 5.4. Comparison with state-of-the-art

**UP-3D**: We compare with two state-of-the-art direct prediction approaches by Lassner *et al.* [22] and Tan *et al.* [43]. We do not include the SMPLify method [7] since a version of this algorithm was used to generate the ground truth for this dataset, so we observed that many estimated reconstructions had only minimal differences from the ground truth. For [22] we use the publicly available code to generate predictions. The complete results are presented in Table 2. Our approach outperforms the other two baselines by significant margins. It is interesting to note that a version of [43], which uses over 100k images (most of them synthetic) with ground truth pose and shape parameters to directly supervise the network (line 'Direct') is outperformed by our approach which does not have access to this data. Finally, in Figure 3, we provide a qualitative comparison with our closest competitor, the direct prediction approach of [22].

**SURREAL**: We compare with two state-of-the-art approaches, one based on iterative optimization, SMPLify [7], and one based on direct prediction [22]. We use the publicly available code for both approaches to generate predictions. For our approach, we train the *PosePrior* using CMU data which we found to be more general than UP-3D. Also, we train two *ShapePriors*, for female and male subjects respectively, since the gender is known for this dataset. We emphasize that the testing was conducted on the Human3.6M part of the dataset to avoid any overlap with the training of the different methods (in terms of images or priors). The complete results are presented in Table 3. Since Lassner *et al.* [22] provide only a non gender-specific model for shape, we also report results considering only the pose estimates, and assuming known shape parameters. Our approach outperforms the other two baselines. For this dataset we observed that because of the challenging color images (low illumination, out-of-context backgrounds, etc), the 2D detections where more noisy than usual, providing some hard failures for the iterative optimization approach [7]. In contrast, our approach was more resistant to these noisy cases recovering a coherent 3D shape in most cases.

|  | Avg |
|---|---|
| Lassner *et al.* [22] (GT shape) | 200.5 |
| Bogo *et al.* [7] (GT shape) | 177.2 |
| Ours (GT shape) | **151.5** |
| Bogo *et al.* [7] | 202.0 |
| Ours | **155.5** |

Table 3. Detailed results on the Human3.6M part of SURREAL [51]. Numbers are mean per vertex errors (mm). "GT shape" indicates that the shape coefficients are known.

|  | Avg |
|---|---|
| Akhter & Black [2]* | 181.1 |
| Ramakrishna *et al.* [35]* | 157.3 |
| Zhou *et al.* [56]* | 106.7 |
| Bogo *et al.* [7] | 82.3 |
| Lassner *et al.* [22] (direct prediction) | 93.9 |
| Lassner *et al.* [22] (optimization) | 80.7 |
| Ours | **75.9** |

Table 4. Detailed results on Human3.6M [17]. Numbers are reconstruction errors (mm). The numbers are taken from the respective papers, except for (*), which were obtained from [7].

**Human3.6M**: Finally, for Human3.6M we evaluate only the estimated 3D pose, since there is no body shape ground truth available. Our network is the same as before (*Priors* trained on CMU), although, we use the 3D joints error for supervision (equation 2), since the focus is on pose. Among others, we compare with the SMPLify method [7] and the direct prediction approach of Lassner *et al.* [22]. Similarly to the other approaches we compare with, we *do not* use any data from this dataset for training. The detailed results are presented in Table 4. Our approach again outperforms the other baselines. Some works have reported better results re-

|  | FB Seg. |  | Part Seg. |  |
|---|---|---|---|---|
|  | acc. | f1 | acc. | f1 |
| SMPLify | 91.89 | 88.07 | 87.71 | 63.98 |
| SMPLify + our anchor | **92.17** | **88.38** | **88.24** | **64.62** |
| SMPLify on GT | 92.17 | 88.23 | 88.82 | 67.03 |

Table 5. Accuracy and f1 scores for foreground-background and six-part segmentation on LSP test set for different versions of SMPLify. Using our direct prediction as an anchor improves vanilla SMPLify, while also achieving a *3x speedup*. The numbers for the first and third rows are taken from [22].

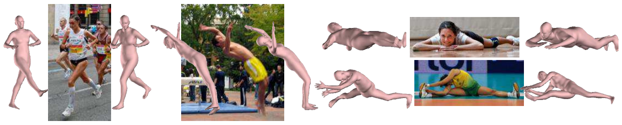

Figure 5. LSP examples with improved SMPLify fits (right side of each image) when our direct prediction is used as an initialization and anchor for the iterative optimization.

sults on Human3.6M (e.g., [27, 31]), but they do so only by leveraging the training data of this dataset for training.

### 5.5. Boosting SMPLify

In the previous section, we validated that our direct prediction approach can achieve state-of-the-art results with a single prediction step. However, we aspire our method to have greater applicability, by being complementary to iterative optimization solutions. In fact, here we demonstrate that our direct predictions can be a useful initialization and provide a reliable anchor for the SMPLify approach [7].

To keep it simple, we make only minor modifications to the SMPLify optimization. First, we use our predicted pose as an initialization, instead of the typical mean pose. Additionally, we avoid the hierarchical four-step optimization, and we limit the whole procedure in a single step. The reason for the multi-stage optimization is to explore the pose space and get a roughly correct pose estimate. However, using our predicted pose as initialization makes this search unnecessary, so we require only the last step of the previously complex optimization scheme. Finally, we add one more data term to the optimization: $E_{anchor}(\boldsymbol{\theta}) = \sum_i \rho(\theta_i - \theta_i^{init})$, to avoid deviations from our predicted, anchor pose. Similarly to [7], we use the Geman-McClure penalty function, $\rho$ [13], for the optimization. This anchoring, does not typically have effect on the quality of the output, but it can accelerate the convergence. We can also use the shape parameters as anchor, but we observed that pose had greater effect than shape on the optimization.

For our evaluation, we use the public implementation of SMPLify and we run the original code, as well as our anchored version, on the LSP test set. The anchored version is *three times faster* on average than vanilla SMPLify. More importantly, this speedup comes also with a quantitative performance *benefit*. In Table 5 we present the segmentation accuracy of different SMPLify versions, by projecting the 3D shape estimate on the image. To demonstrate that the performance benefit of our anchored version is non-trivial, we report the results for running SMPLify on the ground truth 2D joints and silhouettes. Improved fits from the anchored version are presented in Figure 5. These results validate the additional benefit of our direct prediction approach, since it can also enhance current pipelines that rely on iterative optimization.

### 5.6. Running time

Our approach requires a single forward pass from the ConvNet to estimate the full body 3D human pose and shape. This translates to only 50ms on a Titan X GPU. In comparison, SMPLify [7] report roughly 1 minute for the optimization, while the publicly available (unoptimized) code runs on 3 minutes per image on average. When the number of landmarks increases to 91, Lassner *et al.* [22] report that the SMPLify optimization can get two times slower. This makes our direct prediction approach more than three orders of magnitude faster than the state-of-the-art iterative optimization approaches. Regarding other direct prediction approaches, Lassner *et al.* [22] reports runtime of 378ms, but we demonstrate significantly better performance with our end-to-end framework.

## 6. Summary

The goal of this paper was to present a viable ConvNet-based approach to predict 3D human pose and shape from a single color image. A central part of our solution was the incorporation of a body shape model, SMPL, in the end-to-end framework. Through this inclusion we enabled: a) prediction of the parameters from 2D keypoints and silhouettes, b) generation of the full body 3D mesh at training time using supervision for the surface with a per-vertex loss, and c) integration of a differentiable renderer for further end-to-end refinement using 2D annotations. Our approach achieved state-of-the-art results on relevant benchmarks, outperforming previous direct prediction and optimization-based solutions for 3D pose and shape prediction. Finally, considering the efficiency of our approach, we demonstrated its potential to accelerate *and* improve typical iterative optimization pipelines.

**Project Page:** `https://www.seas.upenn.edu/~pavlakos/projects/humanshape`

**Acknowledgements:** We gratefully appreciate support through the following grants: NSF-IIP-1439681 (I/UCRC), ARL RCTA W911NF-10-2-0016, ONR N00014-17-1-2093, DARPA FLA program and NSF/IUCRC.

## References

[1] CMU Graphics Lab Motion Capture Database. 4, 6

[2] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *CVPR*, 2015. 7

[3] T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, and M. Magnor. Optical flow-based 3D human motion estimation from monocular video. In *German Conference on Pattern Recognition*, 2017. 1, 3

[4] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1, 2, 3, 4, 5, 6

[5] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 408–416, 2005. 2, 3

[6] A. O. Bălan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *CVPR*, 2007. 4

[7] 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 *ECCV*, 2016. 1, 3, 4, 5, 6, 7, 8

[8] C.-H. Chen and D. Ramanan. 3D human pose estimation = 2D pose estimation + matching. In *CVPR*, 2017. 2

[9] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross. HS-Nets: Estimating human body shape from silhouettes with convolutional neural networks. In *3DV*, 2016. 2

[10] E. Dibra, H. Jain, C. Oztireli, R. Ziegler, and M. Gross. Human shape from silhouettes using generative HKS descriptors and cross-modal neural networks. In *CVPR*, 2017. 2, 4

[11] E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. Shape from selfies: Human body shape estimation using cca regression forests. In *ECCV*, 2016. 2, 4

[12] G. Gallego and A. Yezzi. A compact formula for the derivative of a 3-D rotation in exponential coordinates. *Journal of Mathematical Imaging and Vision*, 51(3):378–384, 2015. 6

[13] S. Geman and D. McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 1987. 8

[14] P. Guan, A. Weiss, A. O. Bălan, and M. J. Black. Estimating human shape and pose from a single image. In *ICCV*, 2009. 1, 3

[15] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision computing*, 1(1):5–20, 1983. 1

[16] Y. Huang, F. Bogo, C. Classner, A. Kanazawa, P. V. Gehler, I. Akhter, and M. J. Black. Towards accurate markerless human shape and pose estimation over time. In *3DV*, 2017. 1, 3, 4

[17] 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. 4, 5, 6, 7

[18] E. Jahangiri and A. L. Yuille. Generating multiple hypotheses for human 3D pose consistent with 2D joint detections. In *ICCVW*, 2017. 2

[19] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 1, 2, 3, 4, 5, 6

[20] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 6

[21] I. Kokkinos. UberNet: Training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *CVPR*, 2016. 4

[22] 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 *CVPR*, 2017. 1, 3, 4, 5, 6, 7, 8

[23] S. Li and A. B. Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, 2014. 2

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3, 4

[25] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 1, 2, 3

[26] M. M. Loper and M. J. Black. OpenDR: An approximate differentiable renderer. In *ECCV*, 2014. 5

[27] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 2, 4, 8

[28] D. Mehta, H. Rhodin, D. Casas, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017. 2

[29] F. Moreno-Noguer. 3D human pose estimation from a single image via distance matrix regression. In *CVPR*, 2017. 2

[30] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 4

[31] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017. 2, 8

[32] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015. 4

[33] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 3

[34] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2D and 3D human sensing. In *CVPR*, 2017. 2, 4

[35] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3D human pose from 2D image landmarks. In *ECCV*, 2012. 7

[36] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *ECCV*, 2016. 1, 3

[37] G. Riegler, A. O. Ulusoys, and A. Geiger. Octnet: Learning deep 3D representations at high resolutions. In *CVPR*, 2017. 1

[38] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, and S. Fleming. Civilian american and european surface anthropometry resource (caesar), final report. Technical report,

Tech. Rep. AFRL-HE- WP-TR-2002-0169, US Air Force Research Laboratory, 2002. 3, 4, 6

[39] G. Rogez and C. Schmid. MoCap-guided data augmentation for 3D pose estimation in the wild. In *NIPS*, 2016. 2

[40] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net: Localization-classification-regression for human pose. In *CVPR*, 2017. 2

[41] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, 2008. 1, 3

[42] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *ICCV*, 2017. 2

[43] J. K. V. Tan, , I. Budvytis, and R. Cipolla. Indirect deep structured learning for 3D human body shape and pose prediction. In *BMVC*, 2017. 3, 5, 6, 7

[44] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In *ICCV*, 2017. 1

[45] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3D human pose with deep neural networks. In *BMVC*, 2016. 2

[46] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua. Learning to fuse 2D and 3D image cues for monocular body pose estimation. In *ICCV*, 2017. 2

[47] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017. 3

[48] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *CVPR*, 2017. 2

[49] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 4

[50] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 3

[51] G. Varol, J. Romero, X. Martin, N. Mahmood, M. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017. 4, 5, 6, 7

[52] T. von Marcard, B. Rosenhahn, M. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3D human pose estimation from sparse imus. In *Eurographics*, 2017. 1, 3

[53] A. Weiss, D. Hirshberg, and M. J. Black. Home 3D body scans from noisy image and range data. In *ICCV*, 2011. 1, 3

[54] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. MonoPerfCap: Human performance capture from monocular video. *arXiv preprint arXiv:1708.02136*, 2017. 1, 3

[55] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3D human pose estimation in the wild: A weakly-supervised approach. In *ICCV*, 2017. 2

[56] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis. Sparse representation for 3D shape estimation: A convex relaxation approach. *PAMI*, 2016. 7

[57] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. In *CVPR*, 2016. 2