SPEC: Seeing People in the Wild with an Estimated Camera

Muhammed Kocabas[1,2] Chun-Hao P. Huang[1] Joachim Tesch[1] Lea Müller[1]
Otmar Hilliges[2] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]ETH Zurich
{mkocabas,chuang2,jtesch,lea.mueller,black}@tue.mpg.de otmar.hilliges@inf.ethz.ch

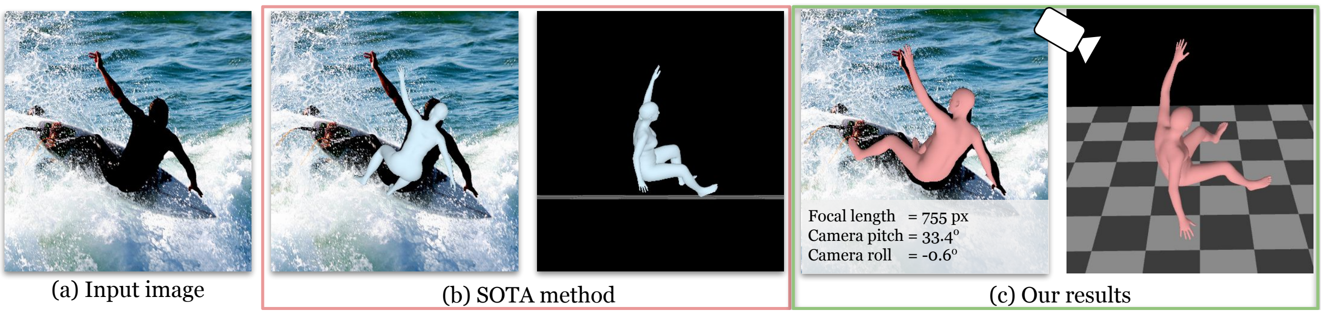


Figure 1: State-of-the-art methods for 3D human pose and shape estimation from images (such as HMR trained with EFT [26] data) struggle with imagery containing perspective effects. (b) In part, this is due to the use of the standard weak perspective camera. (c) SPEC learns to estimate perspective camera parameters and uses these to regress more accurate 3D poses.

Abstract

Due to the lack of camera parameter information for in-the-wild images, existing 3D human pose and shape (HPS) estimation methods make several simplifying assumptions: weak-perspective projection, large constant focal length, and zero camera rotation. These assumptions often do not hold and we show, quantitatively and qualitatively, that they cause errors in the reconstructed 3D shape and pose. To address this, we introduce SPEC, the first in-the-wild 3D HPS method that estimates the perspective camera from a single image and employs this to reconstruct 3D human bodies more accurately. First, we train a neural network to estimate the field of view, camera pitch, and roll given an input image. We employ novel losses that improve the calibration accuracy over previous work. We then train a novel network that concatenates the camera calibration to the image features and uses these together to regress 3D body shape and pose. SPEC is more accurate than the prior art on the standard benchmark (3DPW) as well as two new datasets with more challenging camera views and varying focal lengths. Specifically, we create a new photorealistic synthetic dataset (SPEC-SYN) with ground truth 3D bodies and a novel in-the-wild dataset (SPEC-MTP) with calibration and high-quality reference bodies. Code and datasets are available for research purposes at https://spec.is.tue.mpg.de/.

1. Introduction

Estimating 3D human pose and shape (HPS) from a single RGB image is a core challenge in computer vision and has many applications in robotics, computer graphics, and AR/VR. Reconstructing a high dimensional 3D structure from 2D observations is ill-posed by nature. To overcome this, much attention has been given to structured prediction [16, 20, 62] and incorporating shape and pose priors [47, 69] to guide estimation. Weakly-supervised training of HPS regressors leverages 2D-pose datasets [1, 25, 39] and requires various forms of regularization [28, 35, 71]. Data for full 3D supervision often relies on controlled lab settings [21, 58], synthetic images [61], or, more recently, in-the-wild capture of reference data [43, 63].

Despite rapid progress, we observe that most state-of-the-art (SOTA) methods [5, 7, 15, 24, 26, 28, 29, 32, 35, 36, 47, 52, 59, 73, 74] make several simplifying assumptions about the image formation process itself. First, they all apply a weak perspective or orthographic projection assumption; resulting in a simplified camera model with only three parameters which capture the camera translation relative to the body. Moreover, some [32, 35, 47] set the focal length to a predefined large constant for every input image. Finally, they all assume zero camera rotation, which entangles body rotation and camera rotation, making it extremely hard to correctly estimate the body orientation in 3D. These assumptions are valid for images where bodies are roughly perpendicular to the principal axis and are located far away from the camera. However, in most real world images of people, perspective effects are clearly evident, e.g. foreshortening in selfies. Ignoring perspective projection leads to errors in pose, shape, and global orientation (see Fig. 1).

To overcome these limitations in existing methods, we present *SPEC (Seeing People in the wild with Estimated Cameras)*, the first 3D human pose and shape estimation framework that leverages cues present in the image to extract perspective camera information and exploits this to better reconstruct 3D human bodies from images in the wild. SPEC consists of two parts: camera calibration and body reconstruction. We make contributions to each.

One might hope that embedded EXIF information would be sufficient to address this problem. However, many images lack EXIF information, some applications strip this off, and even if present, converting the stored focal length in millimeters to pixels requires knowing specifics of the image sensor. Given the huge variety of cameras on the market, exploiting this is a non-trivial task. Furthermore, this does not give information about the camera rotation.

Instead, we estimate the camera directly from the RGB image. Recent work [19, 30, 65, 77] casts this ill-posed regression problem as a classification task. However, training such methods with their losses, e.g. cross-entropy and KL-divergence, ignores the natural notion of distance or ordering of the original target space. To address this, we propose a new loss, Softargmax-$\mathcal{L}_2$, to preserve distance during loss calculation. Moreover, we observe that HPS accuracy is quite sensitive to underestimation of focal length and less sensitive to overestimates as also noted by [31, 72]. Therefore, we modify Softargmax-$\mathcal{L}_2$ to be asymmetric such that less penalty is applied when the focal length is overestimated. These novel losses help us to train a better regressor for direct camera calibration, which we term *CamCalib*.

We integrate the regressed camera parameters into two 3D-body-reconstruction paradigms: (1) an optimization-based approach, SMPLify-X [47], and (2) a regression-based one similar to HMR or SPIN [28, 35]. Since SMPLify-X estimates a 3D body my minimizing the difference between projected 3D joints and observed 2D joints, improving the the projective geometry improves the estimated body.

In the case of direct HPS regression from pixels, the estimated camera is employed in two ways: (1) in the reprojection loss similar to the one in SMPLify-X and (2) as *conditioning* for the network by appending the camera parameters to the CNN image features. This second contribution is a key novelty of SPEC, which enables us to disentangle camera and body orientation. SOTA methods [32, 33, 35, 73], cannot do this because the body is estimated in camera space, entangling body orientation and camera rotation.

Training such a body regressor requires in-the-wild images annotated with both 3D human bodies and the camera parameters. Since existing 3D human body datasets [21, 42, 63] contain little variation in camera parameters, we create two new datasets with rich camera variety. First, we create a photorealistic synthetic dataset which has accurate ground-truth human and camera annotations (SPEC-SYN) using ideas from [46]. This dataset is used both for testing and training. Second, we collect a crowdsourced dataset following the Mimic-The-Pose framework [45] (SPEC-MTP). We ask web participants to calibrate their camera and take videos from different angles while mimicking a predefined pose. Then, we obtain pseudo ground-truth labels by fitting the SMPL model to the provided videos while exploiting the predefined pose as a prior. Through extensive experiments and analysis using these new datasets, alongside an existing in-the-wild dataset (3DPW [63]), we show that going beyond the weak-perspective/orthographic assumption improves human pose and shape estimation results.

In summary, our contributions are: (1) We propose a single-view, camera-aware, 3D human body estimation framework that estimates perspective camera parameters from in-the-wild images directly and reconstructs the 3D body without relying on weak-perspective assumptions or offline calibration. (2) We train a neural network to regress the perspective camera parameters given one RGB image, using two novel losses: Softargmax-$\mathcal{L}_2$ and the asymmetric variant to improve the calibration accuracy. (3) Using the estimated camera parameters helps to reconstruct a better 3D body with the optimization-based SMPLify-X algorithm. (4) Conditioning on camera information helps a direct regression approach based on HMR [28] learn to regress better poses. (5) We present two different datasets with ground-truth camera and human body parameters: (i) a photorealistic synthetic dataset, SPEC-SYN, and (ii) a crowdsourced dataset, SPEC-MTP.

2. Related Work

We review work that captures/reconstructs 3D humans under calibrated-camera settings, then focus on our goal: in-the-wild 3D human body reconstruction from monocular RGB. We discuss how prior art simplifies the camera model to make the problem tractable and also discuss relevant methods of camera calibration from one RGB image.

3D HPS Estimation with Calibrated Cameras. To capture human motions in 3D, early work exploits calibrated and synchronized multi-camera settings. They can be largely classified to "bottom up" approaches [4, 6, 9, 14, 57, 76] that assemble 3D body poses from multi-view image evidence (keypoints, silhouettes), and "top down" approaches [3, 11, 12, 20, 27, 62] that deform a pre-defined 3D human template according to detected image features in each view. Powered by CNNs, learning-based methods

[18, 23, 49, 50, 60] gain robustness by training keypoint detection and multi-view pose reconstruction end-to-end.

Some monocular approaches [42, 43, 48, 56] are trained with direct supervision from multi-view data, while others [16, 22, 34, 51] enforce multi-view consistency as weak, or self, supervision. In either way, known intrinsic or extrinsic parameters are always assumed. Yu et al. [72] propose perspective crop layers, which crop the image around a person according to camera parameters and image location, effectively removing some effect of camera geometry.

All these methods require offline calibration and have a risk of overfitting to the cameras used in training and are typically limited to controlled settings.

Single-view HPS Estimation with Unknown Cameras. In early work, Liebowitz and Carlsson [38] use the repetitive structure of a moving person as a cue for camera calibration. Since then, numerous methods reconstruct 3D human bodies given single-view images or videos in uncontrolled settings. Closely related to structure-from-motion and bundle adjustment, [2, 10, 37, 67] take videos as input and jointly estimate cameras and reconstruct human bodies; [17, 40] further ground the bodies in 3D scenes.

We focus on the more general scenario in which the input is a single image. SOTA methods use parametric body models [27, 41, 47, 70] and estimate the parameters either by fitting to detected image features [5, 47, 66] or by regressing directly from pixels with deep neural networks [7, 15, 24, 26, 28, 35, 52, 53, 59, 71, 73, 74]. All these approaches, including non-parametric approaches [36, 54, 55, 75], assume weak perspective/orthographic projection or pre-define the focal length as a large constant for all images. Additionally, they all assume zero camera rotation, which entangles body rotation and camera rotation. As a result, these camera models have only three parameters, capturing the camera translation relative to the body.

Kissos et al. [31] identify this problem and show that replacing focal length with a constant closer to ground truth, i.e. $f = 5000 \rightarrow 2200$, improves results. Wang et al. [64] demonstrate that jointly estimating camera viewpoints and 3D human poses improves cross-dataset generalization. To show this, they train a 3D pose estimation model on available 3D human pose datasets in a supervised way. However, these datasets are limited in terms of camera viewpoint and focal length diversity, background, and number of subjects.

In contrast to the above methods, SPEC generalizes to in-the-wild settings, varied camera intrinsics and viewpoints.

Single-image Camera Calibration. Recent work [19, 30, 65, 77] directly estimates camera parameters from a single image. Zhu et al. [77] also recover the height of some scene objects, e.g. people and cars, together with the camera geometry. They estimate 2D human poses but not 3D bodies. To estimate camera rotations and fields of view, these methods train a neural network to leverage geometric cues in the image without calibration boards or body keypoints. They discretize the continuous space of rotation into bins, casting the problem as a classification task and applying cross-entropy [65] or KL-divergence [19, 77] losses. These losses unfortunately ignore the ordering in target spaces. We devise new losses to retain the concept of distance in the original space, leading to better estimated cameras.

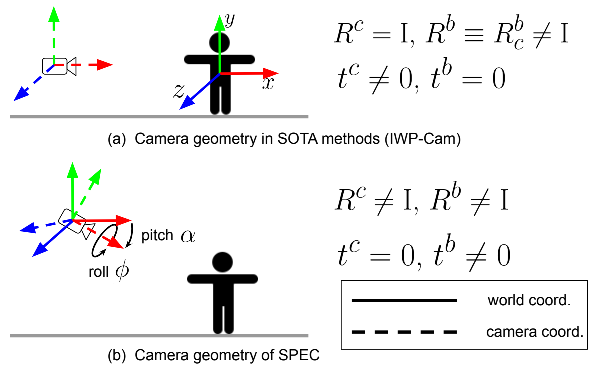


Figure 2: Illustration of IWP-cam and SPEC. R^c and t^c are camera rotation and translation. R^b and t^b are body orientation and translation. All are defined in world coordinate.

3. Method

3.1. Preliminaries

A pinhole camera maps a 3D point $X \in \mathbb{R}^3$ to an image pixel $x \in \mathbb{R}^2$ through $x = K(R^c X + t^c)$, where $K \in \mathbb{R}^{3\times3}$ is the intrinsic matrix storing focal length f_x, f_y and the principal point (o_x, o_y). We follow previous work and omit skew, radial and tangential distortion. Extrinsic parameters are $R^c \in SO(3)$ and $t^c = (t^c_x, t^c_y, t^c_z) \in \mathbb{R}^3$, representing camera rotation and translation in the world coordinate frame, respectively.

We estimate a parameteric human body model, SMPL [41, 47], that deforms a predefined human surface $\mathcal{M}$ according to body pose θ and shape β. When both body translation t^b and body orientation R^b are zero, the mesh is located near the origin of world coordinates, facing the z^+ direction and y^+ is the up-vector, as visualized in Fig. 2(a).

Existing approaches [5, 7, 15, 24, 28, 32, 47, 59, 73] assume zero camera rotation, $R^c = \mathrm{I}$, and estimate the camera translation t^c in two ways: (1) by fitting the body joint coordinates to 2D keypoints [5, 47, 74] or (2) by predicting weak perspective camera parameters (s, t^c_x, t^c_y) with a neural network, where the scale parameter s is converted to t^c_z [28, 29, 32, 35]. See Sup. Mat. for details of this conversion. The underlying assumption here is weak perspective projection. It is assumed that the camera is placed very far from the person, such that depth variations in the z coordinate of the person are negligible compared to the distance from the camera. This is violated regularly in natural images where it is common that the distance from the camera

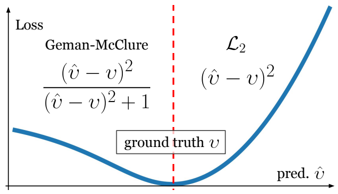


Figure 3: Softargmax-biased-$\mathcal{L}_2$ penalizes underestimates of vfov less than over estimates.

to the body is no more than the height of the body itself.

For intrinsic parameters, [5, 32, 35, 47] set the focal length as a large constant $f_x = f_y = f \equiv 5000$ to meet the weak-perspective assumption and set the principal point (o_x, o_y) at the center of the resized cropped image around the person, while [28, 73, 74] directly apply weak perspective projection $x = sX + t^c$. Despite the differences in modeling projection and translation, one common feature of these simplified cameras is that they only have three unknowns: (s, t^c_x, t^c_y) or equivalently (t^c_x, t^c_y, t^c_z). We refer to them collectively as *IWP-cam* in this paper, standing for Identity rotation and Weak Perspective.

Note that both camera variables (R^c, t^c) and body variables (R^b, t^b) are expressed in world coordinates, not in camera space. Given just one single-view image, the network/optimizer can change both (R^c, t^c) and (R^b, t^b) to explain the image observations. IWP-cam addresses this by assuming $R^c = \mathrm{I}$ and $t^b = 0$ to solve only for camera translation t^c and body orientation R^b, or more precisely, the body orientation in the camera space: $R^b_c = R^c R^b$. See Fig. 2(a). This approach is a key reason why most methods evaluate accuracy after Procrustes alignment of the estimated body to the ground truth.

IWP-cam works well when the weak-perspective assumption holds. However, images captured by cameras with significant pitch and smaller focal lengths, such as those in Fig. 1 and Fig. 5, have foreshortening distortion that breaks this assumption because changes in the z coordinate of the 3D body are no longer negligible compared to the distance from the camera. IWP-cam expects HPS methods to absorb this camera pitch α into the relative body orientation R^b_c, but in practice, due to the mismatch in focal length, the optimization often unnecessarily changes body pose; e.g. the wrong arm and leg poses in Fig. 1(b).

3.2. Camera Calibration from a Single Image

Inspired by single-view camera calibration and metrology [19, 30, 65, 77], we estimate camera rotation R^c and focal length f from a single RGB image. By lifting the zero camera rotation constraint, i.e. $R^c \neq \mathrm{I}$, and directly estimating R^c, we bypass the camera-relative body orientation R^b_c, and thus disentangle camera rotations from body orientations. Doing so allows us to handle perspective/foreshortening distortion while keeping a consistent xz-plane aligned ground plane located at $[0, y, 0]$. Utilizing better focal length also improves the pose estimation quality by leveraging accurate perspective projection.

Specifically, camera rotation is parameterized by three angles: pitch α, roll ϕ and yaw. Since focal length in pixels has an unbounded range and it changes whenever one resizes images, we estimate vertical field of view (vfov) υ in radians and convert it to focal length f_y via:

$$f_y = \frac{\frac{1}{2}h}{\tan(\frac{1}{2}\upsilon)}, \tag{1}$$

where h is the image height in pixels. We follow [77] to assume zero camera yaw and $f_x = f_y = f$. Our camera thus has three more parameters – pitch, α, roll, ϕ, and vertical field of view, υ, in addition to the original (s, t^c_x, t^c_y). Since the three new parameters are all in radians, we choose to learn them with one camera calibration model termed *CamCalib*. We place the camera at the origin so $t^c = [0, 0, 0]$, as depicted in Fig. 2(b), and leave the estimation of body translation t^b to the downstream body estimator.

Many human body reconstruction networks take only a cropped image patch around the person as input. In contrast, CamCalib takes the uncropped full-frame image to predict pitch α, roll ϕ, and vfov υ, which are the same for all subjects in the image. We argue that this is beneficial because the full image contains rich cues that facilitate camera calibration. For example, there are abundant geometric cues, e.g. vanishing points and lines, available to help determine camera rotations and field of view in the original image. Following [19, 30, 77], we use a CNN as the backbone for CamCalib and discretize the spaces of pitch α, roll ϕ, and vfov υ into B bins, converting the regression problem into a B-way classification problem. However, instead of cross-entropy [65] or KL-divergence [19, 77] losses, we aggregate the predicted probability mass using a softargmax operation, i.e. computing the expectation value of the prediction, and measure its difference to the ground truth with an $\mathcal{L}_2$ loss which we term Softargmax-$\mathcal{L}_2$. Thus, we avoid the difficulty of regressing in a continuous target space while retaining the notion of distance in the loss. The detailed formulation of Softargmax-$\mathcal{L}_2$ is provided in the Sup. Mat.

Furthermore, as pointed out by [31, 72] and shown in the Sup. Mat., predicting larger focal lengths than the ground truth (or equivalently smaller fov) is less harmful to the reconstructed 3D poses than predicting smaller focal lengths (larger fov). We therefore apply an asymmetric loss for vfov υ. As shown in Fig. 3, predictions $\hat{\upsilon}$ larger than the ground truth υ yield higher penalty through a standard $\mathcal{L}_2$ loss, while the penalty for smaller predictions $\hat{\upsilon}$ saturates via a Geman-McClure function [13]. We verify the benefits of all these design choices in Sec. 4.

3.3. Optimization approach: SMPLify-X-cam

Next, we showcase how using the estimated camera parameters help human body estimation in an optimization-

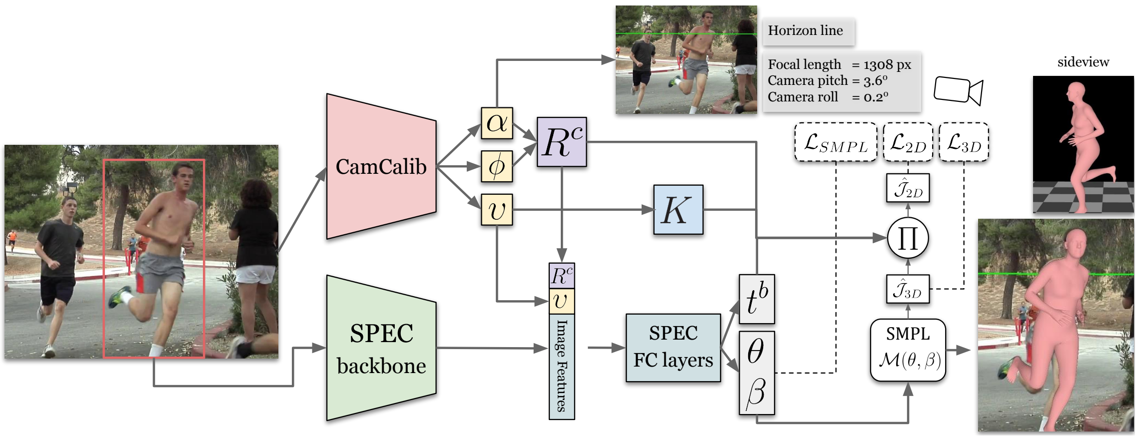


Figure 4: **SPEC overview.** CamCalib takes the whole input image as input and predicts camera pitch α, roll ϕ, and vertical field of view υ. These parameters are then used to construct camera rotation R^c and intrinsics K. Horizon line (green) is rendered following [77] to indicate camera rotations. SPEC takes a cropped bounding box as input and extracts image features using a CNN backbone. Predicted camera parameters from CamCalib are concatenated with image features to estimate SMPL body parameters θ, β along with the body translation t^b. Camera parameters are also taken into account when computing a loss between the projected 3D joints $\hat{\mathcal{J}}_{2D}$ and ground truth.

based approach. To this end, we modify the SMPLify-X method [47]. Given an image, CamCalib predicts camera pitch α, roll ϕ, and vfov υ. We convert them into camera rotation $R^c = R(\alpha)R(\phi)$ and intrinsics, K, storing $f = f_x = f_y$ and the principal point $(o_x, o_y) = (w/2, h/2)$, where f_y is computed from Eq. 1 and w, h are the image width and height in pixels. Then, we estimate the 2D keypoints $\mathcal{J}_{2D}$ using an off-the-shelf 2D keypoint detector [8] and define the the SMPLify-X-cam energy function as:

$$E(\beta, \theta, R^c, K, t^b) = E_J + E_\theta + E_\beta, \quad (2)$$

where β, θ are SMPL shape and pose parameters, t^b is SMPL body translation, E_θ and E_β are pose and shape prior terms, and E_J is the data term. We modify the original SMPLify-X method to take perspective camera parameters into account in the data-term E_J. We obtain SMPL 3D joint locations using a pretrained joint regressor W by $\hat{\mathcal{J}}_{3D} = W\mathcal{M}(\theta, \beta)$. E_J measures the difference between the detected $\mathcal{J}_{2D}$ and the estimated $\hat{\mathcal{J}}_{3D}$, projected on the image by the estimated camera parameters:

$$E_J = \left|\left|\Pi \hat{\mathcal{J}}_{3D} - \mathcal{J}_{2D}\right|\right|_2^2, \text{ where } \Pi = K[R^c| - t^b]. \quad (3)$$

3.4. Learning-based approach: SPEC

To evaluate the effect of estimated camera parameters on a regression-based method, we take a simple and widely-used method, HMR [28], as a backbone, which employs a 2D reprojection loss during training that uses the estimated IWP-cam. We incorporate our estimated camera parameters in two ways: (1) by using them as R^c and K during the projection of 3D joints like in Eq. 3 and (2) by conditioning the fully connected layers, which estimate SMPL parameters, with R^c and υ. Figure 4 gives an overview of SPEC.

Given an image, we first estimate the camera pitch α, roll ϕ, and vfov υ using CamCalib and then convert them into R^c and K as explained in Sec. 3.3. For human body regression, we take a cropped bounding-box image as input and extract image features using a backbone CNN. These image features are concatenated with R^c and υ and fed to an iterative regressor [28] to regress SMPL pose θ and shape β along with body translation t^b. By doing so, SPEC learns to disentangle the SMPL body's global orientation R^b from the camera rotation R^c. See Sup. Mat. for the details of t^b. Then, we obtain SMPL 3D joint locations $\hat{\mathcal{J}}_{3D} = W\mathcal{M}(\theta, \beta)$ and 2D projection $\hat{\mathcal{J}}_{2D} = \Pi\hat{\mathcal{J}}_{3D}$ as in Eq. 3. Overall, our total loss for each training sample is: $\mathcal{L} = \lambda_{3D}\mathcal{L}_{3D} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{SMPL}\mathcal{L}_{SMPL}$ where

$$\mathcal{L}_{3D} = \|\hat{\mathcal{J}}_{3D} - \mathcal{J}_{3D}\|_F^2, \quad (4)$$
$$\mathcal{L}_{2D} = \|\hat{\mathcal{J}}_{2D} - \mathcal{J}_{2D}\|_F^2, \quad (5)$$
$$\mathcal{L}_{SMPL} = \|\hat{\theta} - \theta\|_2^2 + \|\hat{\beta} - \beta\|_2^2, \quad (6)$$

where $\hat{}$ represents the prediction for the corresponding variable. λ's are scalar coefficients to balance the loss terms.

Note that we define both 3D and 2D losses on the body joints. This is because the 3D ground truth we use for training is not always reliable and thus the 2D joints provide important additional image cues, which can be exploited particularly well when using the correct camera geometry.

4. Experiments

We focus the evaluation on CamCalib and SPEC; for evaluation of SMPLify-X-cam, see Sup. Mat.

4.1. Datasets

Pano360 dataset. Previous work [19, 77], uses the SUN360 [68] dataset to train camera calibration networks, which is unfortunately no longer available due to licensing issues. Therefore, we have curated a new dataset of equirectangular panorama images called Pano360. The Pano360 dataset consists of 35K panoramic images of which 34K are from Flickr and 1K rendered from photorealistic 3D scenes. Following previous work [19, 77], we randomly sample the camera pitch, roll, yaw, and vertical field of view to generate 400K training and 15K validation images. We use these to train our CamCalib model.

SPEC-SYN. Existing datasets, used to train HPS regressors, contain limited camera variation. Hence, they are not ideal for training and evaluating the effects of camera estimation on HPS. Therefore, we created a photorealistic synthetic dataset, inspired by AGORA [46], to train and evaluate our model. It has high-quality textured human 3D scans and provides reference SMPL(-X) parameters for them. We place these scans in 5 different large high-quality photorealistic 3D scenes, enabling the generation of many unique views. We randomly sample camera viewpoints, $\alpha \sim \mathcal{U}(-30°, 15°)$ and $\phi \sim \mathcal{N}(0°, 2.8°)$, and focal lengths, $\upsilon \sim \mathcal{U}(70°, 130°)$, to add diversity. In total, we generate 22191 images with 71982 ground truth bodies for training, and 3783 images with 12071 bodies for testing.

SPEC-MTP. To evaluate calibrated HPS (CHPS) methods on real data, we collect a new dataset with high-quality pseudo ground truth using Amazon Mechanical Turk (AMT). Following the idea of the MTP dataset [45], we ask AMT workers to mimic 10 poses for which we have 3D ground truth. While the person maintains a pose, a second person records a video from different viewpoints. In addition, the worker calibrates the camera and provides their height and weight. We extend SMPLify-XC [45] and use the calibrated camera to fit the SMPL-X model to multiple video frames. See Sup. Mat. for details. In total, we collect 64 videos of 7 subjects (4 male, 3 female) and extract 3284 images at a frame rate of 1 fps.

Other datasets. To train 3D CHPS estimation, we use the 3DPW [63], COCO [39], MPI-INF-3DHP [42], and Human3.6M [21] datasets. We evaluate SPEC using separate test data: 3DPW-test, SPEC-SYN, and SPEC-MTP. Since there are no ground truth 3D body and camera annotations for COCO, we use CamCalib to estimate camera parameters and SMPLify-X-cam to obtain pseudo 3D body ground truth, using EFT [26] annotations as the initialization.

4.2. Evaluation metrics

The mean per joint position error (MPJPE), Procrustes-aligned mean per joint position error (PA-MPJPE), and per vertex error (PVE) are the most commonly-used evaluation metrics in the literature. PA-MPJPE exists as a metric specifically because current HPS methods that use IWP-cam reconstruct bodies in camera coordinates $\hat{\mathcal{J}}_{3D}^{cam}$. Procrustes alignment "hides many sins" in that it removes the rotation of the body caused by an unknown camera pose. See Sup. Mat. for details of how PA-MPJPE and MPJPE are computed.

Methods	**vfov** $\upsilon° \downarrow$	**pitch** $\alpha° \downarrow$	**roll** $\phi° \downarrow$
ScaleNet [77]	5.68	2.61	1.41
CamCalib (KL Loss)	3.53	2.32	1.15
CamCalib (Softargmax-$\mathcal{L}_2$)	3.34	2.06	1.11
CamCalib (Softargmax-biased-$\mathcal{L}_2$)	**3.24**	**1.94**	**1.02**

Table 1: **Regressing camera parameters.** CamCalib methods are trained and tested on the Pano360 dataset. ScaleNet [77] results use the authors' implementation.

Instead, we propose variants of MPJPE and PVE that compute the error in world coordinates without the need of camera information and dub them W-MPJPE and W-PVE. Since SPEC disentangles camera and body rotations, the predictions reside in world coordinates $\hat{\mathcal{J}}_{3D}^{world}$ and W-MPJPE is computed as $\|\hat{\mathcal{J}}_{3D}^{world} - \mathcal{J}_{3D}\|$. For existing SOTA methods, we report two versions of W-MPJPE: (1) $\|\hat{\mathcal{J}}_{3D}^{cam} - \mathcal{J}_{3D}\|$ and, (2) $\|R^{c^{-1}} \hat{\mathcal{J}}_{3D}^{cam} - \mathcal{J}_{3D}\|$, where R^c is the estimated camera rotation by CamCalib. By reporting (2), we do not assume known camera rotations for any method and compare them all in world coordinates. This also illustrates the effect of using CamCalib with prior work. We discuss these metrics in greater detail and report MPJPE & PVE in the Sup. Mat.

4.3. Implementation details

CamCalib. We follow the implementation of [19, 77] but use ResNet-50 as the backbone and predict pitch α, roll ϕ, and vfov υ with separate fully-connected layers. Each parameter has $B = 256$ bins and we apply Softargmax-biased-$\mathcal{L}_2$ for υ, Softargmax-$\mathcal{L}_2$ for α and ϕ. The model is trained with images of varied resolutions for 30 epochs. The Pano360 dataset is used for training and evaluation.

SPEC. Similar to the original HMR [28], we use a ResNet-50 backbone, followed by fully-connected layers that iteratively regress SMPL parameters. We do not apply HMR's adversarial discriminator since we use psuedo-ground-truth 3D training data. The Adam optimizer with a learning rate of $5e^{-5}$ is used. For the first 150 training epochs we use the COCO and SPEC-SYN datasets, and then incorporate MPI-INF-3DHP and Human3.6M. Total training takes around 175 epochs, ~4-5 days.

Note that CamCalib and SPEC are trained separately. Training them jointly is not possible because we lack an *in-the-wild* dataset that has both ground-truth bodies and diverse camera focal lengths and views. SPEC-MTP meets these requirements but is small so we us it for evaluation. During inference, CamCalib and SPEC run jointly.

4.4. Single-image camera calibration results

Table 1 reports the mean angular error in camera pitch α, roll ϕ, and vfov υ for different camera calibration methods on the Pano360 test set. For reference, we inlcude the open-source implementation of ScaleNet [77], which uses a different backbone than CamCalib. We also train ScaleNet with our backbone on Pano360 (CamCalib (KL loss) in Table 1). We evaluate the effect of different loss functions, i.e. KL-divergence, Softargmax-$\mathcal{L}_2$, and Softargmax-biased-$\mathcal{L}_2$, and define the final CamCalib network to be the best performing version (Softargmax-biased-$\mathcal{L}_2$).

4.5. SPEC evaluation

Tables 2, 3, and 4 show the results of recent SOTA methods on the SPEC-MTP, SPEC-SYN and 3DPW datasets.

The correct metric. We believe W-MPJPE is the metric that best reflects performance in real-world applications, so we report W-MPJPE, PA-MPJPE, and W-PVE. MPJPE is also reported and discussed in Sup. Mat. For W-MPJPE and W-PVE, we report both definitions from Sec. 4.2, i.e. (1)/(2) in the tables. Note that SPEC has the same error under both metrics. PA-MPJPE has only one entry because Procrustes alignment removes the effect of camera rotation (and more); this effectively hides the fact that SOTA methods do not estimate global pose well.

Comparison to the state-of-the-art. To compute the performance of SOTA methods, we use their open source implementations. We use HMR* as our IWP-cam baseline, which is HMR [28] trained with the same datasets as SPEC, i.e. COCO, SPEC-SYN, MPI-INF-3DHP, and Human3.6M. Again, we do not use HMR's discriminator since we train with ground-truth or pseudo ground-truth 3D labels. For I2L-MeshNet [44], we use the SMPL output of this method instead of non-parametric mesh to be able to report W-PVE and denote this with †.

Since W-MPJPE and W-PVE measure the error w.r.t. the body in world coordinates, these measures reveal the performance improvement of SPEC over the SOTA when the camera deviates from the IWP-cam assumption. Compared a dataset like 3DPW, SPEC-MTP and SPEC-SYN have significantly more variety in focal lengths and viewpoints as shown in Fig. 5. As a result, SPEC yields a larger improvement in W-MPJPE and W-PVE over the SOTA for these datasets. Using explicit camera information is a key driver of this improvement. The improvement in PA-MPJPE is less significant, suggesting that the largest improvements come from estimating the body in world coordinates rather than better articulated pose. This is valuable in many applications, e.g. human-scene interaction, where bodies and objects are often reconstructed from distinct methods but should reside in a common space.

Figure 6 analyzes W-MPJPE on SPEC-SYN for different camera viewpoints and focal lengths. SPEC results are similar across different camera settings, while the HMR* is less robust to values that deviate from its assumptions.

Methods	W-MPJPE	PA-MPJPE	W-PVE
GraphCMR [36]	175.1 / 166.1	94.3	205.5 / 197.3
SPIN [35]	143.8 / 143.6	79.1	165.2 / 165.3
PartialHumans [52]	158.9 / 157.6	98.7	190.1 / 188.9
I2L-MeshNet† [44]	167.2 / 167.0	99.2	199.0 / 198.1
HMR* [28]	142.5 / 128.8	**71.8**	164.6 / 150.7
SPEC	**124.3 / 124.3**	**71.8**	**147.1 / 147.1**

Table 2: Results of SOTA methods on SPEC-MTP dataset. We use the implementations provided by the authors to obtain results. HMR* means that we train HMR using the same data as SPEC for fair comparison. †means we use the SMPL output of this method instead of the non-parametric mesh to be able to report W-PVE. All numbers are in *mm*.

Methods	W-MPJPE	PA-MPJPE	W-PVE
GraphCMR [36]	181.7 / 181.5	86.6	219.8 / 218.3
SPIN [35]	165.8 / 161.4	79.5	194.1 / 188.0
PartialHumans [52]	169.3 / 174.1	88.2	207.6 / 210.4
I2L-MeshNet† [44]	169.8 / 163.3	82.0	203.2 / 195.9
HMR* [28]	128.7 / 96.4	55.9	144.2 / 111.8
SPEC	**74.9 / 74.9**	**54.5**	**90.5 / 90.5**

Table 3: Results of SOTA methods on SPEC-SYN. See Table 2 caption.

Methods	W-MPJPE	PA-MPJPE	W-PVE
GraphCMR [36]	137.8 / 129.4	69.1	158.4 / 152.1
SPIN [35]	122.2 / 116.6	59.0	140.9 / 135.8
Partial Humans [52]	139.4 / 132.9	76.9	160.1 / 152.7
I2L-MeshNet† [44]	133.3 / 119.6	60.0	154.5 / 141.2
HMR* [28]	119.2 / **104.0**	53.7	136.2 / **120.6**
SPEC	**106.4** / 106.4	**53.2**	**127.4** / 127.4

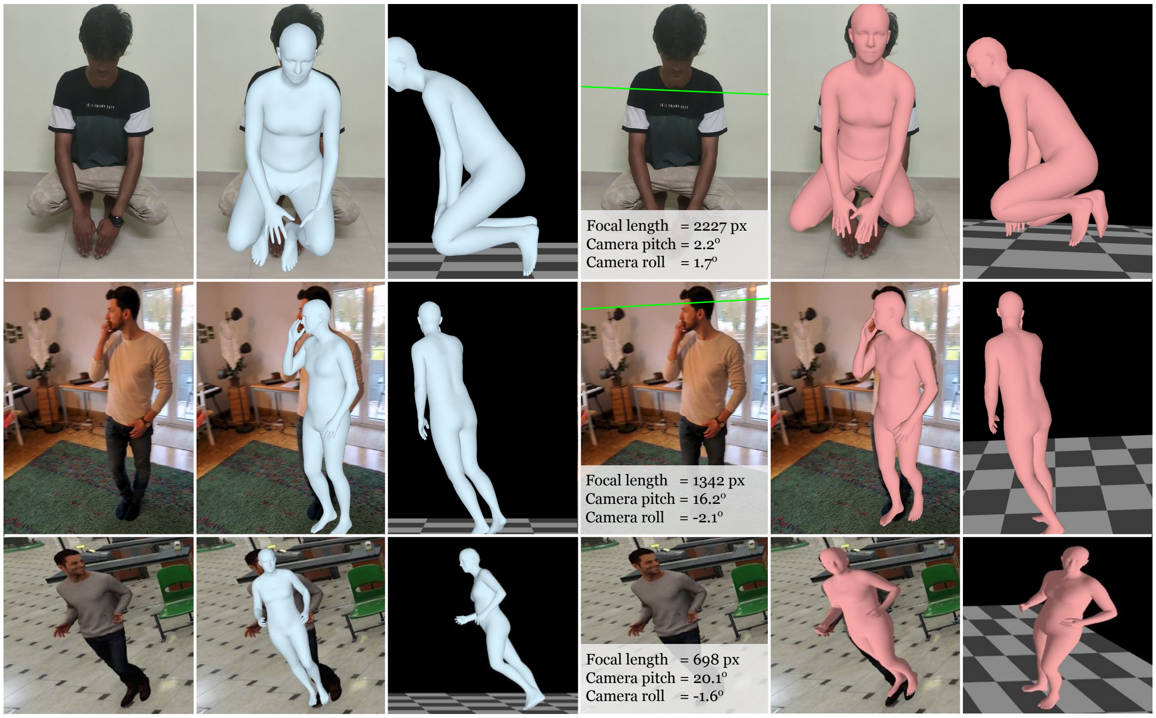
Table 4: Results of SOTA methods on 3DPW test set. See Table 2 caption.

Methods	W-MPJPE	PA-MPJPE	W-PVE
HMR*	128.7 / 96.4	55.9	144.2 / 111.8
HMR* + c	120.4 / 84.2	54.0	135.3 / 98.8
HMR* + c + f	118.3 / 85.1	54.0	132.8 / 99.7
HMR* + c + f + R^c	77.2 / 77.2	55.3	93.8 / 93.8
SPEC	**74.9 / 74.9**	**54.5**	**90.5 / 90.5**

Table 5: Ablation studies with SPEC-SYN. c: using the image center as camera center. f and R^c: using CamCalib estimated focal length and camera rotation, respectively.

Ablation experiments. We ablate different camera parameters and components of SPEC to investigate their effect on performance. We report results on the SPEC-SYN test data (Table 5) because it has challenging cameras; for ablation results on 3DPW please see Sup. Mat.

We use HMR* as our baseline and use the same datasets and training configurations for all methods. To obtain the predicted 2D joints, $\hat{\mathcal{J}}_{2D}$, HMR* uses the bounding-box



(a) Input Image (b) HMR* - front (c) HMR* - side (d) CamCalib (e) SPEC - front (e) SPEC - side

Figure 5: Qualitative results. Top & middle: SPEC-MTP; bottom: SPEC-SYN. We also provide failure cases in Sup. Mat.

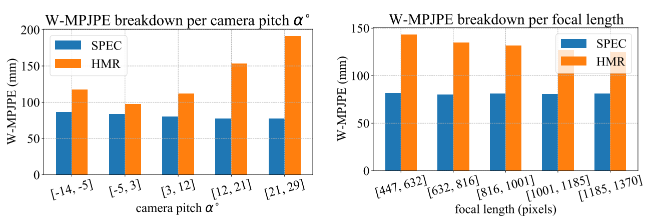


Figure 6: Breakdown of W-MPJPE per camera focal length and pitch range.

center as the principal point. We start by changing it to the image center, i.e. $o_x = w/2, o_y = h/2$ denoted as "HMR* + c". This ensures a better projective geometry than using the bounding-box center as the image center and already improves results. Next, we replace the fixed focal length of 5000 with the values estimated by CamCalib, denoted as "HMR* + c + f". "HMR* + c + f + R^c" uses R^c instead of the identity matrix as the camera rotation during projection. Finally, SPEC uses c, f, and R^c both during projection and as a conditioning input to the final stage of HMR* predictions. Overall, improving the camera model improves W-MPJPE and W-PVE. Conditioning the network on the camera parameters helps, but we suspect that better camera conditioning schemes can be employed to make the network more aware of the camera geometry.

Qualitative results. Figure 5 shows representative results. HMR* assumes IWP-cam and yields incorrect body poses (legs in top row) and incoherent body orientations (middle & bottom); SPEC predicts overall more globally coherent bodies as can be seen in side-view images.

5. Conclusion

In this paper, we demonstrate that a) camera geometry can be estimated from images and b) can effectively be leveraged to improve 3D HPS accuracy. Existing methods make simplifying assumptions about the camera: weak-perspective projection, large constant focal length, and zero camera rotation. To go beyond these simple assumptions, we introduce SPEC, the first 3D HPS method that regresses a perspective camera from a single image and employs this to reconstruct 3D human bodies more accurately. Using the estimated camera parameters improves both SOTA camera regression methods and HPS regression methods. We introduce two new datasets, i.e. SPEC-MTP and SPEC-SYN, with accurate camera and 3D body annotations to showcase the effect of SPEC through ablation studies and comparison with the SOTA and to foster future research in this area.

Acknowledgements: We thank Emre Aksan, Shashank Tripathi, Vassilis Choutas, Yao Feng, Priyanka Patel, Nikos Athanasiou, Yinghao Huang, Cornelia Kohler, Hongwei Yi, Dimitris Tzionas, Nitin Saini, and all Perceiving Systems department members for their feedback and the fruitful discussions. This research was partially supported by the Max Planck ETH Center for Learning Systems.

Disclosure: https://files.is.tue.mpg.de/black/CoI/ICCV2021.txt

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[3] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007. 2

[4] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*, 2016. 1, 3, 4

[6] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2

[7] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, 2020. 1, 3

[8] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 5

[9] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[10] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *European Conference on Computer Vision*, 2020. 3

[11] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[12] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2

[13] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, LII-4:5–21, 1987. 4

[14] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3D structure with a statistical image-based shape model. In *International Conference on Computer Vision*, 2003. 2

[15] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 1, 3

[16] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. DeepCap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2020. 1, 3

[17] Nils Hasler, Bodo Rosenhahn, Thorsten Thormahlen, Michael Wand, Jürgen Gall, and Hans-Peter Seidel. Markerless motion capture with unsynchronized moving cameras. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 224–231. IEEE, 2009. 3

[18] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[19] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matthew Fisher, Emiliano Gambaretto, Sunil Hadap, and Jean-François Lalonde. A perceptual measure for deep single image camera calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4, 6

[20] Chun-Hao Huang, Benjamin Allain, Edmond Boyer, Jean-Sebastien Franco, Federico Tombari, Nassir Navab, and Slobodan Ilic. Tracking-by-detection of 3d human shapes: from surfaces to volumes. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2017. 1, 2

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2014. 1, 2, 6

[22] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[23] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision*, 2019. 3

[24] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3

[25] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1

[26] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020. 1, 3, 6

[27] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2, 3

[28] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and

pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 4, 5, 6, 7
[29] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3
[30] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2, 3, 4
[31] Imry Kissos, Lior Fritz, Matan Goldman, Omer Meir, Eduard Oks, and Mark Kliger. Beyond weak perspective for monocular 3d human pose estimation. In Adrien Bartoli and Andrea Fusiello, editors, *European Conference on Computer Vision*, pages 541–554, Cham, 2020. Springer International Publishing. 2, 3, 4
[32] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 1, 2, 3, 4
[33] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision*. IEEE, Oct. 2021. 2
[34] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3D human pose using multi-view geometry. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 3
[35] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision*, 2019. 1, 2, 3, 4, 7
[36] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3, 7
[37] Vincent Leroy, Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, and Grégory Rogez. SMPLy benchmarking 3d human pose estimation in the wild. In *International Conference on 3D Vision, 3DV*, 2020. 3
[38] David Liebowitz and Stefan Carlsson. Uncalibrated motion capture exploiting articulated structure constraints. volume 51, pages 171–187. Springer, 2003. 3
[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 1, 6
[40] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M Rehg, and Siyu Tang. 4d human body capture from egocentric video via 3d scene grounding. *arXiv preprint arXiv:2011.13341*, 2020. 3
[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015. 3
[42] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *International Conference on 3DVision*, 2017. 2, 3, 6
[43] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *International Conference on 3DVision*, 2018. 1, 3
[44] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, 2020. 7
[45] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self contact and human pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 6
[46] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 6
[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3, 4, 5
[48] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3
[49] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *International Conference on Computer Vision*, 2019. 3
[50] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3
[51] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3
[52] Chris Rockwell and David F. Fouhey. Full-body awareness from partial observations. In *European Conference on Computer Vision*, 2020. 1, 3, 7
[53] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. In *arXiv*, 2020. 3
[54] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision*, 2019. 3
[55] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3
[56] Yeji Shen and C-C Jay Kuo. Multi-view matching (mvm): Facilitating multi-person 3d pose estimation learning with

action-frozen people video. In *arXiv preprint*, 2020. 3

[57] Leonid Sigal, Alexandru Balan, and Michael J Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in Neural Information Processing*, 2008. 2

[58] 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*, 87(1):4–27, Mar. 2010. 1

[59] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision*, 2020. 1, 3

[60] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision*, 2020. 3

[61] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[62] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *SIGGRAPH*, pages 1–9. 2008. 1, 2

[63] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision*, 2018. 1, 2, 6

[64] Zhe Wang, Daeyun Shin, and Charless C Fowlkes. Predicting camera viewpoint improves cross-dataset generalization for 3d human pose estimation. In *European Conference on Computer Vision*, 2020. 3

[65] Scott Workman, Menghua Zhai, and Nathan Jacobs. Horizon lines in the wild. In *British Machine Vision Conference*, 2016. 2, 3, 4

[66] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[67] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. MonoClothCap: Towards temporally coherent clothing capture from monocular rgb video. In *International Conference on 3DVision*, pages 322–332, 2020. 3

[68] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 6

[69] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3d human shape and articulated pose models. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 1

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

[71] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3d pose and shape estimation by dense render-and-compare. In *International Conference on Computer Vision*, 2019. 1, 3

[72] Frank Yu, Mathieu Salzmann, Pascal Fua, and Helge Rhodin. PCLs: Geometry-aware neural reconstruction of 3d pose with perspective crop layers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 4

[73] Sun Yu, Bao Qian, Liu Wu, Fu Yili, and Mei Tao. CenterHMR: a bottom-up single-shot method for multi-person 3d mesh recovery from a single image. In *arxiv:2008.12272*, August 2020. 1, 2, 3, 4

[74] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. 2020. 1, 3, 4

[75] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[76] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4D association graph for realtime multi-person motion capture using multiple video cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[77] Rui Zhu, Xingyi Yang, Yannick Hold-Geoffroy, Federico Perazzi, Jonathan Eisenmann, Kalyan Sunkavalli, and Manmohan Chandraker. Single view metrology in the wild. In *European Conference on Computer Vision*, 2020. 2, 3, 4, 5, 6, 7