3D HUMAN POSE ESTIMATION AND TRACKING IN THE WILD

Dorian F. Henning
Department of Computer Science
Imperial College London
London, SW7 2AZ
d.henning@imperial.ac.uk

Alp Guler
Department of Computer Science
Imperial College London
London, SW7 2AZ
r.guler@imperial.ac.uk

Stefan Leutenegger
Department of Computer Science
Imperial College London
London, SW7 2AZ
s.leutenegger@imperial.ac.uk

Stefanos Zafeiriou
Department of Computer Science
Imperial College London
London, SW7 2AZ
s.zafeiriou@imperial.ac.uk

May 17, 2020

ABSTRACT

Human motion estimation in 3D is a fundamental problem in computer vision and robotics related tasks. Current state-of-the-art systems have a high precision particularly for the 2D keypoint re-projection, they work mostly in image space and struggle to locate the human in a global, metric coordinate frame that is needed for robotics applications. In this work, we demonstrate how to extend a model-based human pose and shape estimator to track 3D human keypoints in the wild. We introduce a method to compute a virtual camera with desired properties that allows us to remove focal distortion from image crops and rescale them to a desired size. Furthermore, we show how to utilise a weak perspective camera model and a SLAM system to estimate the metric, absolute position and orientation of the human body. The human pose and position estimates can be used to apply state estimation techniques to filter human motion and movement, independent from any camera ego motion on in-the-wild video sequences.

***K*eywords** Human Pose Estimation · Robotics · Filtering · Human Motion Tracking · Global Coordinates

1 Introduction

Human Pose Estimation in unknown and unstructured environments is a fundamental and challenging problem in computer vision. Reliably estimating 2D or 3D joint locations and body parts is necessary for different applications including, but not limited to robotic vision, action classification, and human-robot interaction.

Traditionally, there have been two major approaches to solve this highly unconstrained problem. Bottom-up approaches use feature localization techniques to determine keypoint locations in 2D or 3D, and group these keypoints to individual subjects in a subsequent step. Top-down approaches are model-fitting techniques, where the unique topology of humans is exploited to fit a kinematic skeleton or human mesh model to a subject.

While bottom-up approaches are fast, reliable, and robust with respect to viewpoint changes and multiple subjects in one image [1, 2, 3], they suffer from a lack of predictive ability when it comes to (self-)occlusion. Particularly in robotic vision and indoor environments, challenging viewpoints (low camera position), partial visibility of subjects (close-up views where parts of the subjects are outside the view), and occlusion by furniture pose major demands at human pose estimation pipelines. Moreover, after localisation of 2D keypoints, it is a challenging task to lift them to 3D using skeleton models or convolutional neural networks alone [4, 5], without the use of depth sensors [6]. Therefore recently, model-based techniques gained a lot of attention, that enforce additional constraints on human pose and shape, while limiting the amount of parameters to estimate.

Parametric human mesh models, for example the Skinned Multi-Person Linear (SMPL) model [7], describe a human mesh with a very limited amount of shape parameters and joint angles. Mesh regression networks use deep learning techniques to regress pose and shape parameters of a mesh model and a weak perspective camera model from a tightly cropped human detection. This approach is very promising regarding predictive capabilities, since the deep learned model takes all pixel values into account, restricts certain unnatural poses and shapes, and can account for (self-)occluded body parts.

Human pose estimation in video sequences acquired from a moving camera, often encountered in robotic vision, poses additional challenges, like association of subjects over time (tracking) and measurement noise (camera ego-motion jitter and detection noise). Classical approaches to filter predicted body positions, poses, and shapes showed to be unsuccessful because of the very few constraints that can be imposed on the system. However, by using state estimation techniques (e.g. SLAM) to "remove"(estimate) the ego motion of the camera, we can show that this enables us to more robustly estimate human poses in 3D space and remove prediction noise.

We propose a 3D human pose estimation framework that identifies humans as mesh model in global world coordinates with a moving camera. With our method it is possible to filter or optimize body shape parameters, and joint and body motion that is isolated from any camera ego-motion to reduce position error and increase smoothness of estimated joint trajectories.

1.1 Contributions

Our work is split into 3 parts which we will describe in the following sections.

In a first part, we introduce a mathematically sound formulation to remap any bounding box in an image with known camera calibration on a realistic image crop that accounts for any perspective and lens distortion. This is necessary, since many human pose estimation frameworks rely on tight image crops of humans as network inputs, which are assumed as quasi-orthographic images. This method is to our knowledge the first reported method that mathematically justifies this approach.

Second, using this virtual pinhole camera model derived from the detection box and predicted scale parameter from the model-based human mesh prediction framework (e.g. [8, 9]), the metric transformation between the predicted body frame and the real pinhole camera frame can be determined. Furthermore, using the absolute position of the real pinhole camera estimated by a visual-inertial SLAM system (e.g. OKVIS [10]), we compute the position of the human in metric, global coordinates which is necessary for the use in robotic applications or high-fidelity reconstruction with a moving camera.

Lastly, we are proposing the use of classical state estimation techniques like Kalman filtering to improve the localization of body center and respective joints in the tracking setting.

2 Related Work

Single Image Human Pose Estimation 3D human pose estimation is usually approached via two ways, either by looking at joint locations only, or by using full body models.

If the human pose is estimated as a point cloud of 3D joints, the most common way is to lift a set of 2D joints into 3D using a convolutional neural network [4, 5] or more traditionally a dictionary of human poses or pose prior [11, 12]. One problem with this approach is, that is does not generalize well to in-the-wild images with previously unseen pose configurations or noisy 2D keypoint detections. While you can train a network to directly regress 3D human pose information using 3D datasets such as MPI-INF-3DHP [13] or Human3.6M [14], this approach does not generalize well to in-the-wild images like the 3DPW dataset [15].

Another, more promising approach is to use human body models such as SCAPE [16] or the newer SMPL [7], which can also capture facial expression and hand articulation [17]. These body models can be fitted to images using either an optimization approach [18, 19, 20], or directly regress body model parameters using a convolutional neural network with supervision of 2D and 3D keypoints [8, 21, 22]. Since the supervision signal is very weak and the available datasets are limited, Kolotouros *et al.*proposed a combination of the optimization and regression approaches with SPIN [9]. In this work, the authors implement a SMPLify optimization routine [18] in the network training loop that fits an initial model to the 2D ground truth keypoints. The optimized body model parameters are used as supervision signals for the regression network.

Despite achieving great results for single image human pose estimation, and are widely transferrable to in-the-wild images, these methods yield jittery results when applied to videos.

Video Human Pose Estimation While there are many approaches to estimate human motion from videos looking only on joint locations [23, 24, 25], we will focus on methods that use parametric human body models like SMPL.

Arnab *et al.*[26] extend the previously introduced SMPLify routine to incorporate temporal information. By posing smoothness constraints for pose parameters and consistent body shape, they essentially formulate the task of human pose estimation in videos as a bundle adjustment problem. Huang *et al.*[27] use multiple views of the same scene and silhouette constraints to improve single frame fitting. Furthermore, a discrete cosine transform prior for human joint trajectories is applied to smooth the body motion. A deep learning based approach was proposed in [28], where the network learns human kinematics by prediction of past and future image frames. Just recently, Kocabas *et al.*[29] introduce both a temporal encoder with a gated recurrent unit (GRU) to estimate temporal consistent motion and shapes, and a training procedure with a motion discriminator that forces the network to generate feasible human motion.

While these results produce visually pleasing results and have prove to have lower reprojection errors than frame based approaches, they fail to estimate metric transformations between body and camera coordinate frames. Furthermore, the deep learned approaches all use an orthographic camera model assumption during the training loop, which works fine with internet videos and common datasets and benchmarks, but fails in cases with close-up images where for example focal distortion plays a significant role.

3 Preliminaries

In the following section, we define some basic notation, the used human mesh model, and the optional detection framework.

3.1 Basic Notation

In this paper, we will use the following notation to describe the different coordinate frames: A reference coordinate frame is denoted as $\underrightarrow{\mathcal{F}}_A$. The homogeneous transformation from $\underrightarrow{\mathcal{F}}_B$ to $\underrightarrow{\mathcal{F}}_A$ is denoted as $\boldsymbol{T}_{AB}$, and consists of the rotation matrix $\mathbf{C}_{AB}$, and translation vector ${}_A\mathbf{r}_{AB}$. The change of coordinate frame of the homogeneous point ${}_B\boldsymbol{r}$ from $\underrightarrow{\mathcal{F}}_B$ to $\underrightarrow{\mathcal{F}}_A$, is computed by matrix multiplication ${}_A\boldsymbol{r} = \boldsymbol{T}_{AB}\,{}_B\boldsymbol{r}$. A point in the RGB image I_C from the pinhole camera C is given as $\mathbf{u}$ and the corresponding homogeneous coordinate is denoted as $\boldsymbol{u}$. The camera calibration or intrinsic matrix $\mathbf{K}_C$ has the form

$$\mathbf{K}_C = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

with f_x, f_y being the focal length in x and y direction, and its principal point $[c_x, c_y] \in \mathbb{R}^2$, and is used to transform a point from homogeneous camera coordinates to image coordinates $\boldsymbol{u} = \mathbf{K}_C\boldsymbol{r}$.

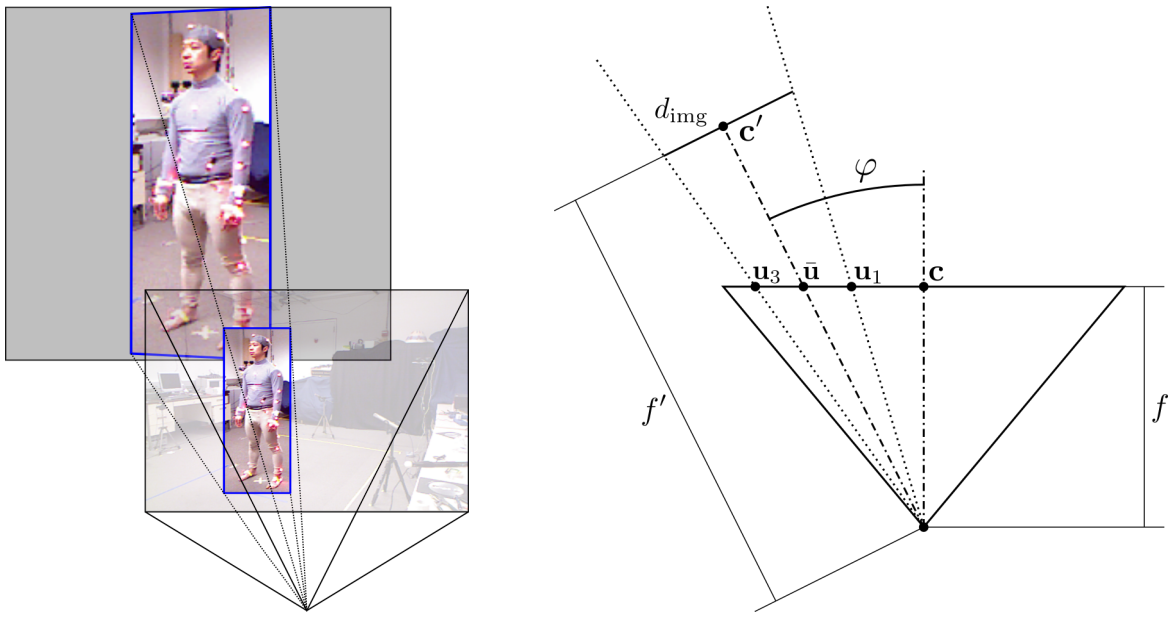
Throughout the next sections, we will denote the following coordinate frames: the inertial world reference frame $\underrightarrow{\mathcal{F}}_W$, the frame of the physical pinhole camera $\underrightarrow{\mathcal{F}}_C$ with z-axis pointing towards the scene, the virtual pinhole camera frame $\underrightarrow{\mathcal{F}}_{C'}$, and the (human) body centric frame $\underrightarrow{\mathcal{F}}_B$.

3.2 Body Model

The SMPL [7] model is a parametric human mesh model, that supplies a function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ of the shape parameters $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ and returns a mesh $\mathbf{M} \in \mathbb{R}^{N\times 3}$ as a set of $N = 6890$ vertices given in body centric frame $\underrightarrow{\mathcal{F}}_B$. The body shape is given as a 10 dimensional vector $\boldsymbol{\beta} \in \mathbb{R}^{10}$, where the components are determined by a principal component analysis, to capture as many variability of human shape as possible. The body pose $\boldsymbol{\theta}$ captures each of the 23 joint angles and the body root orientation in either axis-angle representation with dimensionality $\dim(\boldsymbol{\theta}) = (23 + 1) \times 3 = 72$, or as a 6D rotation-representation as introduced in [30] with $\dim(\boldsymbol{\theta}) = (23 + 1) \times 6 = 144$. The joints can be defined as a linear combination of mesh vertices. For k joints, a linear regressor $\mathbf{W} \in \mathbb{R}^{k\times N}$ can be pre-trained and the 3D body joints are ${}_B\mathbf{X} = \mathbf{W}\,\mathbf{M}$ with ${}_B\mathbf{X} \in \mathbb{R}^{k\times 3}$.

3.3 Human Detection

Human mesh regression networks require a high quality crop of the image to reliably regress the pose and shape parameters as well as body scale and translation. This can be achieved using either bounding boxes derived from 2D joint locations detected by a keypoint detector [1, 2], or any bounding box detector that can detect humans (e.g. YOLO [31], Faster-RCNN [32]). For our evaluation on Berkeley Multimodal Human Action Database (MHAD) [33], we use the provided 2D joint locations.



(a) Conceptual virtual camera in 3D

(b) Conceptual virtual camera in 2D.

Figure 1: Virtual camera models in 3D and in 2D, visualizing the concepts of focal undistortion and remapping (a), and the identification of rotation angle φ and new focal length f_F (b).

4 Methodology

The following section is divided into 5 parts. First, we describe our contribution towards the introduction of a new camera model to properly prepocess images that result in realistic image crops of humans. Then we explain the image rectification and the used human mesh regressor. In the last two parts, we derive how to estimate the metric body position in a stationary, inertial reference frame, and how to use this information in a for movement estimation and filtering.

4.1 Perspective Camera Model

Natural images I taken from a camera are subject to lens and perspective distortion. Most deep networks that predict human shape and pose only use small, human-centered image crops as inputs, where it is sufficient to treat those crops as quasi-orthographic images (images taken with a large focal length compared to any image dimension $f \gg d_{\text{img}}$). Although the results of these networks, in particular human mesh regressors (HMR) are astonishing [8, 9, 29], for robotics application where humans will be close to the camera and not centered in the image, this assumption does not hold. In our proposed pipeline, we introduce a new virtual camera to align the camera center with the center of the bounding box, rescale the image, and furthermore derive accurate positional information between the camera and body reference coordinate frames.

The image cropping and rectification is one central part of our software contribution, and is done in three steps. In the following, we will define C as the real, physical camera used to acquire the images, and C' as a virtual camera, which has some desired properties that allow us to remap the perspective projection I of the scene on to a square image I' of fixed size. We assume to have knowledge of the calibration of camera C, with focal lengths f_x, f_y and principal point $\mathbf{c} = [c_x, c_y]^\mathsf{T}$. Using the bounding box defined by four corners $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4$, starting in the top left and numbered clockwise, camera matrix $\mathbf{K}_C$, and the normalized image size $d_{\text{img}} = 224\,\text{px}$, a rotation matrix $\mathbf{C}_{CC'}$ and a virtual camera matrix $\mathbf{K}_{C'}$ are determined. It is necessary to emphasize the fact that the focal point of the virtual camera coincides with the focal point of the real camera, as demonstrated in Fig. 1b.

The principal point $\mathbf{c}' = [c'_x, c'_y]^\mathsf{T}$ of the virtual camera should be aligned with the center of the bounding box by a rotation in 3D as conceptually shown in 2D in Fig. 1a, so that the subject is centered in the image crop. Furthermore, $\mathbf{c}'$ is defined to lie centered in the image crop I':

$$c'_x = c'_y = \frac{d_{\text{img}}}{2}. \tag{2}$$

A rotation matrix $\mathbf{C}_{CC'}$ can be found via Rodrigues formula, using the principal axis of the real camera $\boldsymbol{c}$ and the center of the bounding box

$$\overline{\boldsymbol{u}} = 0.5(\boldsymbol{u}_1 + \boldsymbol{u}_3) = 0.5(\boldsymbol{u}_2 + \boldsymbol{u}_4), \tag{3}$$

in homogeneous coordinates, as the two vectors that define the plane of rotation. The angle of rotation φ is determined using the inner vector product

$$\varphi = \arccos(\overline{\boldsymbol{u}}^{\mathsf{T}}\boldsymbol{c}). \tag{4}$$

The focal length candidates $f'_{(i)}$ are found by subsequently posing the condition that the rotated backprojection of the corners of the bounding box $\mathbf{C}_{C'C}\ \mathbf{K}_C^{-1}\,\boldsymbol{u}_{(i)}$ should be equal to the backprojection of the corners of the normalized image on to the edges of the image $\mathbf{K}_{C'}^{-1}\,\boldsymbol{u}'_{(i)}$. Unless the we encounter the unlikely case that we have an image centered bounding box with the same aspect ration of our original image, this condition leads to 8 equations that can be analytically solved.

Starting from the following equation:

$$\mathbf{K}_{C'}^{-1}\,\boldsymbol{u}'_{(i)} = \mathbf{C}_{C'C}\,\mathbf{K}_C^{-1}\,\boldsymbol{u}_{(i)}, \tag{5}$$

to find equations for the focal length $f' = f'_x = f'_y$, the inverse camera matrix $\mathbf{K}_{C'}^{-1}$ is decomposed

$$\mathbf{K}_{C'}^{-1} = \frac{1}{f'}\begin{bmatrix} 1 & 0 & -c'_x \\ 0 & 1 & -c'_y \\ 0 & 0 & 1 \end{bmatrix} = \frac{1}{f'}\mathbf{A}, \tag{6}$$

and we define a new matrix $\mathbf{B} \in \mathbb{R}^{3\times 3}$,

$$\mathbf{B} = \mathbf{C}_{C'C}\ \mathbf{K}_C^{-1}. \tag{7}$$

Reformulating (5) with (6) and (7), we arrive at

$$\frac{1}{f'}\mathbf{A}\,\boldsymbol{u}'_{(i)} = \mathbf{B}\,\boldsymbol{u}_{(i)}. \tag{8}$$

We define the row vectors of the respective matrices $\mathbf{A}$ and $\mathbf{B}$ with

$$\mathbf{A} = [\mathbf{a}_x, \mathbf{a}_y, \mathbf{a}_z]^{\mathsf{T}} \in \mathbb{R}^{3\times 3}, \text{ and} \tag{9}$$

$$\mathbf{B} = [\mathbf{b}_x, \mathbf{b}_y, \mathbf{b}_z]^{\mathsf{T}} \in \mathbb{R}^{3\times 3}, \tag{10}$$

respectively.

Then, we insert the respective corner of the bounding box $i \in [1, \ldots, 4]$ and inspect each of the two non-trivial rows $j \in [x, y]$ individually, choosing the smallest candidate to assure that the whole bounding box is reprojected into the normalized virtual image I':

$$f' = \min_{i,j} \frac{\mathbf{b}_j^{\mathsf{T}}\,\boldsymbol{u}_{(i)}}{\mathbf{a}_j^{\mathsf{T}}\,\boldsymbol{u}'_{(i)}}. \tag{11}$$

Intuitively, a larger focal length will cause the projection of the scene to enlarge and therefore some parts would lie outside the image crop.

4.2 Image Rectification

To control the image size and remove any focal distortion that is caused by the bounding box not being aligned with the principal point of the camera, we perform a remapping. This is performed by reprojecting every pixel of the image crop $\boldsymbol{u}'$ on to the original image:

$$\boldsymbol{u} = \mathbf{K}_C\,\mathbf{C}_{CC'}\ \mathbf{K}_{C'}^{-1}\,\boldsymbol{u}', \tag{12}$$

using bilinear interpolation.

This remapping transforms the image to an image crop (around the bounding box) such that the network input looks as if the center of the bounding box is aligned with the principal point of the camera. Furthermore, since we limit any arising focal distortion to a minimum, the assumption of a quasi-orthographic network input image is sufficient, and network inference should perform better in general.

An example of the resulting images after remapping and cropping can be seen in Fig. 1a

4.3 Human Mesh Regressor

The regression network we use was proposed by Kolotouros *et al.*[9] and is based on the work from Kanazawa *et al.*[8], but uses a 6D representation for 3D rotations as introduced by Zhou *et al.*[30]. The human mesh regressor (HMR) approximates a function that provides model (β, θ) and camera $\Pi = [s, \mathbf{t}],\ s \in \mathbb{R}, \mathbf{t} \in \mathbb{R}^2$ parameters as output from a forward pass of an image. Normally, these networks are trained with a reprojection loss on the 2D joint positions. This, however, puts a lot of pressure on the network during training, since it has to search in a high-dimensional space for a suitable solution that explains the joint locations through body shape, pose and camera parameters. The main contribution of SPIN [9] is a special training procedure, where the loss is directly computed on the model and camera parameters.

The optimization procedure introduced by Bogo *et al.*[18] is performed in the training loop of the HMR, initialized by the network output. This routine optimizes for model and camera parameters that allows them to compute a loss directly on the regressed parameters, so the optimizer does not need to search in a high-dimensional parameter space. While this new training procedure increases learning efficiency and the improves the quality of the regressed results, it does not affect inference time in comparison to [8].

4.4 Metric Body Frame Transformation

The outputs of the HMR are body pose and shape parameters (θ, β) and a weak perspective camera model with translation $\mathbf{t} \in \mathbb{R}^2$ and scale parameter s. Since a weak perspective projection follows the same rules as an orthographic projection, except for the scale parameter s, to render the mesh $\mathcal{M}(\beta, \theta)$ into the image crop I_C, a large but arbitrary focal length can be chosen. The only necessary condition is that during training and inference, this focal length remains the same. That is because in the case of both mesh regressors [8, 9], this focal length is used to compute a part of the loss function with which the neural network is trained.

In the following section, we describe our method to compose a homogeneous transformation $\boldsymbol{T}_{CB}$ that allows us to transfer our mesh vertices ${}_B\mathbf{v}_i$ into our camera frame $\underrightarrow{\mathcal{F}}_C$ utilizing our virtual camera model $\mathbf{K}_{C'}$, the corresponding rotation $\mathbf{C}_{CC'}$ and the predicted weak perspective camera model $[s, \mathbf{t}]$.

As a first part, the homogeneous transformation $\boldsymbol{T}_{C'B}$ between the virtual camera reference frame $\underrightarrow{\mathcal{F}}_{C'}$ and the body centric frame $\underrightarrow{\mathcal{F}}_B$ is computed. This transformation is composed of a rotation to align the bode frame to the virtual camera reference frame and a translation. Since the body mesh orientation in the body frame is described in the first element of the pose vector θ (the global orientation of the SMPL mesh) the axis of $\underrightarrow{\mathcal{F}}_B$ and $\underrightarrow{\mathcal{F}}_{C'}$ are already aligned, and the rotation $\mathbf{C}_{C'B}$ describes only a swap of axis. While the translation in x and y is directly predicted with $\mathbf{t}$, translation in z is defined by the weak perspective camera equation with

$$t_z = 2\,\frac{f'}{s\; d_{\text{img}}}. \tag{13}$$

The homogeneous transformation between $\underrightarrow{\mathcal{F}}_C$ and $\underrightarrow{\mathcal{F}}_{C'}$ is a pure rotation with $\mathbf{C}_{CC'}$, as derived in 4.1. The resulting transformation is

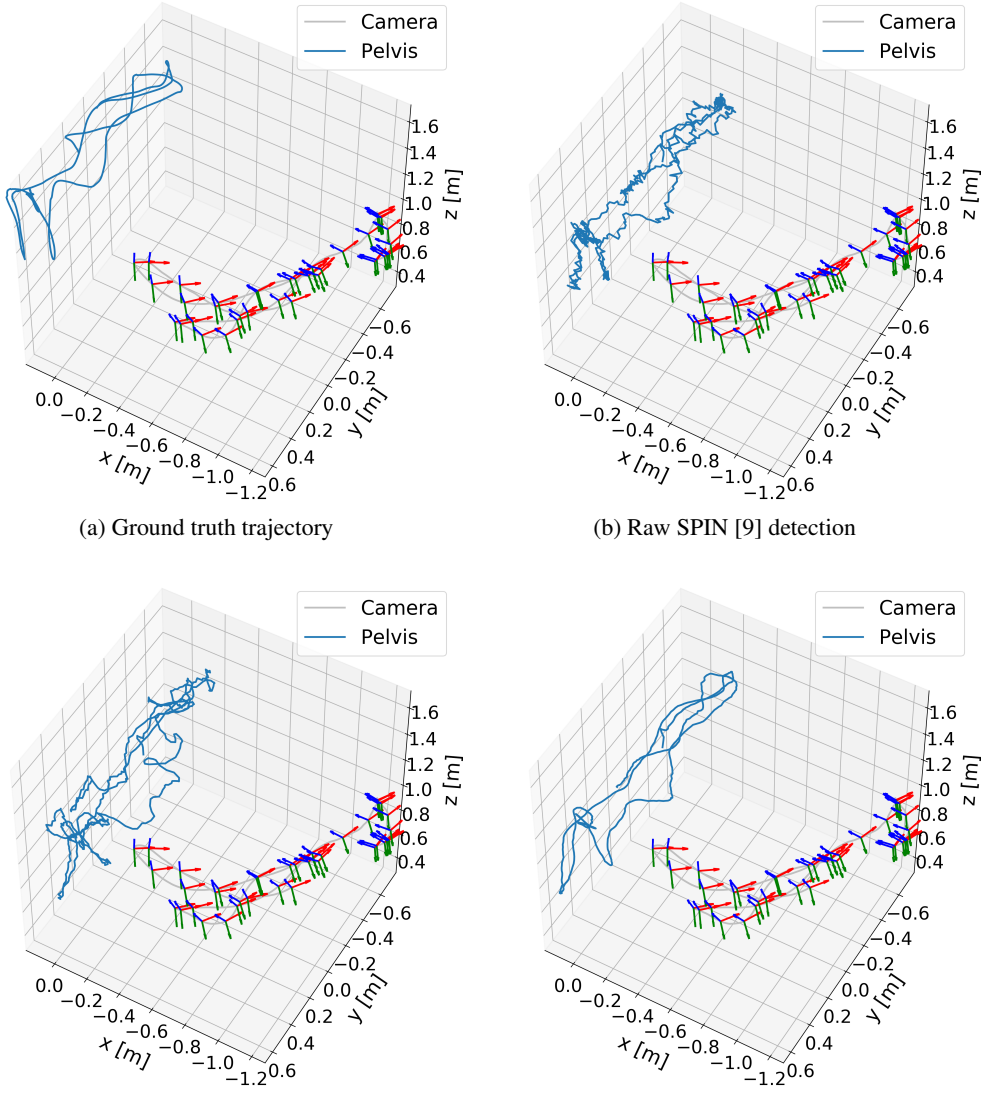
$$\boldsymbol{T}_{CB} = \boldsymbol{T}_{CC'}\,\boldsymbol{T}_{C'B}\,. \tag{14}$$

4.5 3D Human Movement Estimation and Filtering

In this section we explain the contribution of our work regarding human movement estimation. It is important to note the distinction between human motion and movement: while human motion looks at the change in human pose and configuration, we define human movement as the actual body frame displacement with respect to the inertial reference frame.

Intuitively, while state-of-the-art frameworks that look at a time sequences [26, 29] treat joint displacement only in the body reference frame, our work enables us to distinguish between a human walking through a room and a human walking on a treadmill. While this information might be irrelevant for high-fidelity reconstruction tasks, it is essential for robotic applications, where the accurate, relative 3D position between a human and a moving robot is necessary for save motion planning, human-robot interaction, and 3D scene understanding.

Current state-of-the-art methods try to achieve smooth body motion by learning to discriminate between valid and invalid temporal changes of the body pose $[\theta_0, \ldots, \theta_T]$, or by enforcing a consistent body shape $\overline{\beta}$. While this approach achieves visually pleasing results and beats other frame-by-frame approaches such as SMPLify [18] or SPIN [9] by exploiting temporal context to achieve higher accuracy, it only considers human motion in reference to the body centric



(a) Ground truth trajectory

(b) Raw SPIN [9] detection

(c) Filtered trajectory in camera frame

(d) Filtered trajectory in inertial frame (our method)

Figure 2: Pelvis trajectories of subject picking up a box in blue; camera positions and orientations are visualised with a coordinate frame.

coordinate frame. Furthermore, to smooth the relative position between body and camera frame, current methods only filter the position in 2D camera space, which is sufficient for no or only slow camera motion (MPI-INF-3DHP [13], Human 3.6M [14], 3DPW [15]). For fast and jittery camera motion like on a robotic platform, it is necessary to filter or optimize in a reference frame that is decoupled from any camera ego motion.

With the in (14) derived homogeneous transformation, it is now possible to address the mesh vertices $\boldsymbol{v}_i$ and joints $\mathbf{X}$ in an inertial, metric reference frame $\underrightarrow{\mathcal{F}}_W$:

$$ {}_W\boldsymbol{X} = \boldsymbol{T}_{WB}\, {}_B\boldsymbol{X}, \text{ and} \tag{15} $$

$$ {}_W\boldsymbol{v}_i = \boldsymbol{T}_{WB}\, {}_B\boldsymbol{v}_i. \tag{16} $$

4.6 Kalman Filter

In the following part, we will introduce the Kalman Filter that is used to filter the the different states of the human mesh model over time. The three state variables that are used to filter the human shape, motion and movement are

$$ \mathbf{x}_{\beta} := (\beta_1, \ldots, \beta_{10}), \tag{17} $$

$$ \mathbf{x}_{\boldsymbol{\theta}^{(i)}} := (\boldsymbol{\theta}^{(i)}, \dot{\boldsymbol{\theta}}^{(i)}), \tag{18} $$

$$ \mathbf{x}_{\mathbf{r}} := ({}_W\mathbf{r}, {}_W\mathbf{v}\,), \tag{19} $$

with:

- $\beta_1, \ldots, \beta_{10}$: individual shape parameters of the body mesh,
- $\boldsymbol{\theta}^{(i)}$: individual orientation of joint i in 6D rotation representation [30],
- ${}_W\mathbf{r}$: position of the body root (pelvis) expressed in coordinate frame $\underrightarrow{\mathcal{F}}_W$,
- ${}_W\mathbf{v}$: velocity of the body root (pelvis) expressed in coordinate frame $\underrightarrow{\mathcal{F}}_W$.

We arrive at the following differential equations for the state propagation from time step $k-1$ to k:

$$ \dot{\boldsymbol{\beta}}_k = 0, \tag{20} $$

$$ \dot{\boldsymbol{\theta}}_k^{(i)} = \dot{\boldsymbol{\theta}}_{k-1}^{(i)}, \tag{21} $$

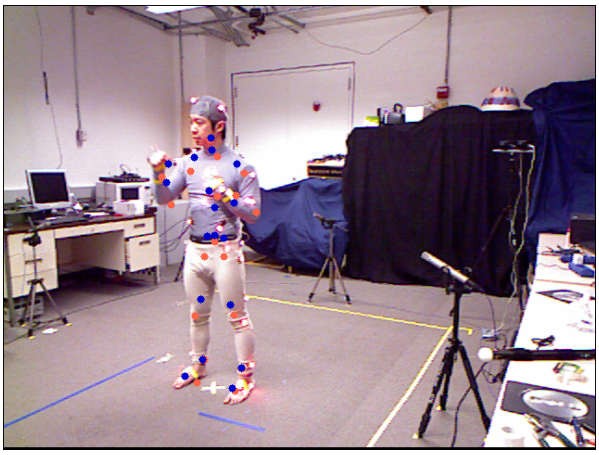
$$ \ddot{\boldsymbol{\theta}}_k^{(i)} = 0, \tag{22} $$

$$ {}_W\dot{\mathbf{r}}_k = {}_W\mathbf{v}_{k-1}\,, \tag{23} $$

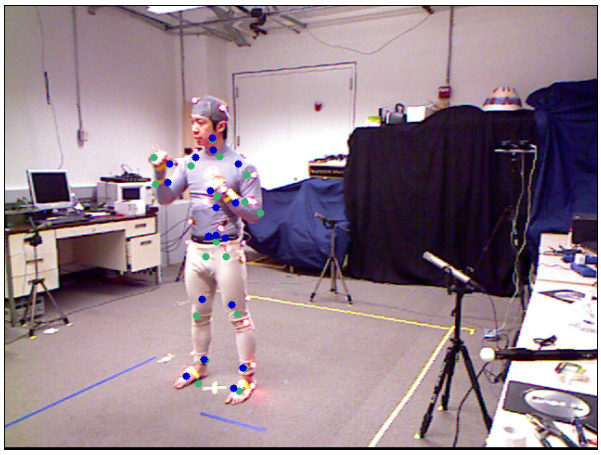
$$ {}_W\dot{\mathbf{v}}_k = 0. \tag{24} $$

We assume to have a constant velocity model of our 3D body root position ${}_W\mathbf{r}$, and constant body shape parameters $\boldsymbol{\beta}$. For the individual joint orientations $\boldsymbol{\theta}^{(i)}$, we also assume to have constant change over time. The observable states are $\boldsymbol{\beta}$, $\boldsymbol{\theta}^{(i)}$, and ${}_W\mathbf{r}$. In the filter update step, we use the measurements of our observable states, $\tilde{\boldsymbol{\beta}}_k$, $\tilde{\boldsymbol{\theta}}_k^{(i)}$, and ${}_W\tilde{\mathbf{r}}_k$, to update our states.

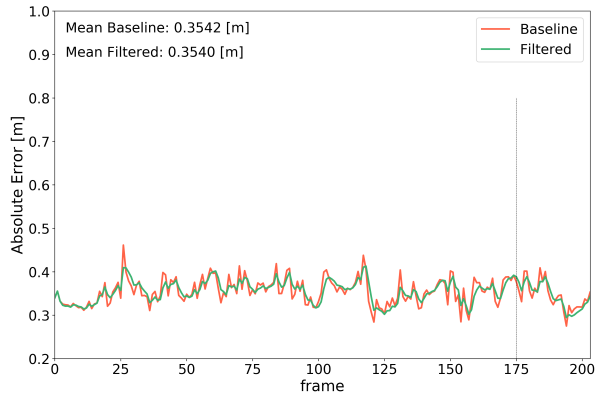
In the following section, we elaborate further assumptions for our filtering step. For measurements $\mathbf{z} \in \mathbb{R}^N$, the covariance matrix of the observation noise $\mathbf{R} \in \mathbb{R}^{N \times N}$ is assumed to be constant and is computed using the measurements of all observable states throughout one clip of the dataset (with ≈ 200 frames). The process noise for a state $\mathbf{x} \in \mathbb{R}^M$ is assumed to be small, and approximated with $\mathbf{Q} = 0.01\, \mathbf{I}_M \in \mathbb{R}^{M \times M}$, with $\mathbf{I}_M$ being the M-dimensional identity matrix. The initial covariance is assumed to be large and the covariance matrix is estimated as $\mathbf{P}_0 = 10\, \mathbf{I}_M \in \mathbb{R}^{M \times M}$. Furthermore, the covariance of all unobservable states is additionally penalized with a factor to account for larger uncertainty.



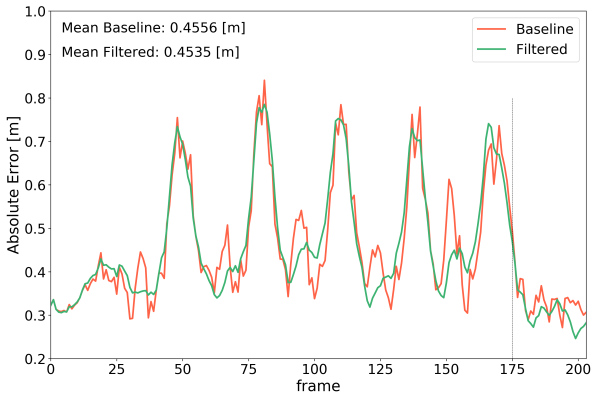
(a) Raw SPIN [9] (red) and ground truth (blue) joints



(b) Filtered (green) and ground truth (blue) joints



(c) Absolute error of pelvis (root) joint; static joint example (joint does not move excessively during clip)



(d) Absolute error of right hand joint; dynamic joint example (subject is boxing with his right and left hand.)

Figure 3: Qualitative and quantitative results on a selected clip from Berkeley MHAD [33] dataset with boxing subject. The vertical line indicates the shown frame above.

5 Results

In this section, we present some qualitative results of our method on the Berkeley MHAD [33] dataset and our own data (denoted as HPE3D). Any evaluation of our method on the 3DPW [15] dataset failed, because the supplied camera pose ground truth only allowed to recover accurate relative poses between the body frame $\underrightarrow{\mathcal{F}}_B$ and camera frame $\underrightarrow{\mathcal{F}}_C$, but no accurate absolute positions.

In Fig. 2, we present in four separate plots from HPE3D data the advantage of our method. The raw detection without any post processing is shown in 2b. When the movement of the human is filtered the camera reference frame as shown in 2c, one can see that the detection noise can not be removed, while additional drift from the camera ego motion is introduced, although the camera motion is slow and smooth. Following our method, the decoupled trajectory of the human can be filtered as well, leading to an improved result shown in 2d. While the trajectory shape in 2d matches the ground truth in 2a, the offset of the predicted position can not be removed by filtering only.

Some qualitative example images from the Berkeley MHAD dataset are shown in Fig. 3. The raw detection without any filtering applied is shown in red, while the filtered results are given in green (for both the whole joint reprojection and absolute error plots). Ground truth 2D joints from the dataset are given in blue (the joint definitions of the Berkeley MHAD dataset did not match the joints supplied by the SMPL model). On top of the slightly reduced absolute error through filtering, one can observe that the filtering does not introduce a strong delay in the joint motion and that even for challenging and fast motions such as boxing as shown in Fig. 3b, the reprojection lies close to the joint ground truth.

6 Conclusion

In this paper we present our contributions towards human pose estimation in the wild. A novel method is introduced to define a virtual camera with intrinsics and orientation that achieve minimal focal distortion when cropping the image to a normalized shape. Furthermore, we propose a procedure to derive a metric transformation between the physical camera and the body frame using an estimated weak perspective camera model and the proposed virtual camera. Lastly, we show how to use classical state estimation techniques that are able to filter human joint trajectories in an inertial reference frame, decoupled from any camera ego motion. The method was validated on the Berkeley MHAD [33] and our own dataset, and some qualitative and quantitative results were shown.

Future work should explore a numerical validation on custom data, since all currently available datasets are either with a stationary camera [33, 13, 14] or only supply correct relative camera positions [15], rendering our contribution redundant. Also, the combination of Additionally, we aim to experiment with mesh fitting using a differentiable renderer (e.g. OpenDR [34]), using depth information from RGB-D datasets.

References

[1] Zhe Cao, Tomas Simon, Shih En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[2] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional Multi-person Pose Estimation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 11 2017.

[3] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient Online Pose Tracking. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2 2018.

[4] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 5 2017.

[5] Ruiqi Zhao, Yan Wang, and Aleix Martinez. A Simple, Fast and Highly-Accurate Algorithm to Recover 3D Shape from 2D Landmarks on a Single Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(12):3059–3066, 9 2016.

[6] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. 3D Human Pose Estimation in RGBD Images for Robotic Task Learning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018.

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, 11 2015.

[8] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end Recovery of Human Shape and Pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.

[10] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *International Journal of Robotics Research (IJRR)*, 2015.

[11] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, 10 2015.

[12] Jack Valmadre and Simon Lucey. Deterministic 3D human pose estimation using rigid structure. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 467–480, 2010.

[13] 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 *Proceedings of the International Conference on 3D Vision (3DV)*, 2017.

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(7):1325–1339, 7 2014.

[15] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 9 2018.

[16] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape Completion and Animation of People. *ACM Transactions on Graphics*, 24(3), 2005.

[17] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 561–578, 10 2016.

[19] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the People: Closing the Loop Between 3D and 2D Human Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4704–4713, 1 2017.

[20] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes The Importance of Multiple Scene Constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[21] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural Body Fitting: Unifying Deep Learning and Model-Based Human Pose and Shape Estimation. In *Proceedings of the International Conference on 3D Vision (3DV)*, 8 2018.

[22] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to Estimate 3D Human Pose and Shape from a Single Color Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 5 2018.

[23] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3D Human Pose from Structure and Motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 679–696. Springer Verlag, 11 2017.

[24] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3D pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 69–86, 11 2017.

[25] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7745–7754. IEEE Computer Society, 11 2018.

[26] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3390–3399, 5 2019.

[27] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards Accurate Marker-less Human Shape and Pose Estimation over Time. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2017.

[28] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D Human Dynamics from Video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12 2019.

[29] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2020.

[30] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12 2019.

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2016.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2017.

[33] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, Rene Vidal, and Ruzena Bajcsy. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, 2013.

[34] Matthew M. Loper and Michael J. Black. OpenDR: An approximate differentiable renderer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–169, 2014.