KAMA: 3D Keypoint Aware Body Mesh Articulation

Umar Iqbal Kevin Xie Yunrong Guo Jan Kautz Pavlo Molchanov
NVIDIA

2021 International Conference on 3D Vision (3DV) | 978-1-6654-2688-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/3DV53792.2021.00078

Abstract

We present KAMA, a 3D Keypoint Aware Mesh Articulation approach that allows us to estimate a human body mesh from the positions of 3D body keypoints. To this end, we learn to estimate 3D positions of 26 body keypoints and propose an analytical solution to articulate a parametric body model, SMPL, via a set of straightforward geometric transformations. Since keypoint estimation directly relies on image clues, our approach offers significantly better alignment to image content when compared to state-of-the-art approaches. Our proposed approach does not require any paired mesh annotations and provides accurate mesh fittings through 3D keypoint regression only. Results on the challenging 3DPW and Human3.6M show that our approach yields state-of-the-art body mesh fittings.

1. Introduction

The estimation of a human body mesh from a single RGB image is of great interest for numerous practical applications. The state-of-the-art methods [12, 17, 27, 30] in this area use deep neural networks with a fully-connected output layer, and directly regress the parameters of a parametric mesh model from the input image. While the performance of these methods has improved significantly, learning a mapping between images and mesh parameters in this way is highly non-linear. Therefore, these methods often suffer from low localization accuracy. Specifically, while these methods estimate parameters that are plausible, the resulting meshes are often misaligned with the visual content, in particular, the wrists and feet regions (See Fig. 1). Additionally, these methods require a large number of images annotated with ground-truth meshes which is very hard to acquire specifically in unconstrained scenes.

On the other hand, recent methods for 3D keypoint regression [23, 68, 70] can accurately localize body keypoints with their 2D projections aligning well with the image content. Instead of learning a direct mapping between input images and 3D coordinates [59, 67, 71, 84], these methods first estimate an intermediate volumetric [43, 55, 68, 70] or heatmap-like [23] representation, and then recover the 3D coordinates from them. This results in better 3D keypoint localization as better correspondences can be built between spatial image locations and the output 3D representation through fully-convolutional neural networks. Motivated by this, some recent works for mesh estimation also encode mesh vertex coordinates in heatmap-like representation [10]. While they show impressive performance, it comes at the cost of requiring large amounts of images annotated with body pose and shape labels.

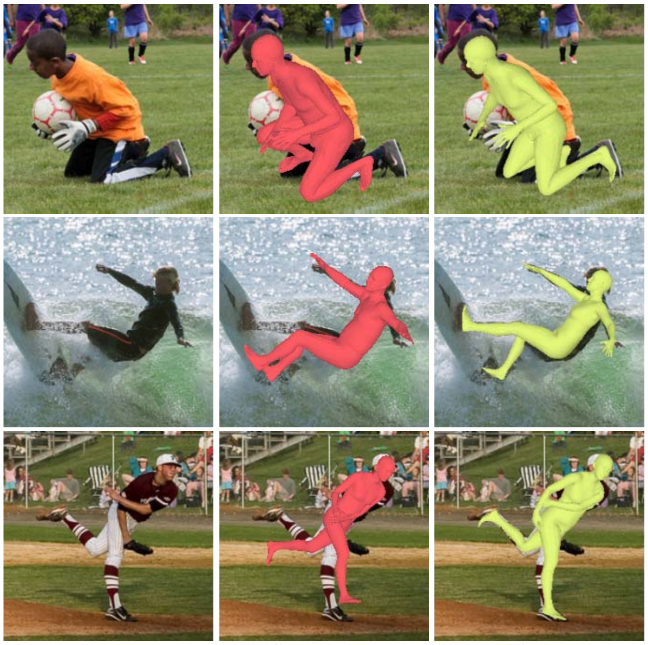

Figure 1. Qualitative comparison with the state-of-the-art method SPIN [30]. While mesh predictions of SPIN [30] are plausible, they do not align well with the image content, especially around the hand and feet regions. In contrast, our method yields accurate meshes with better alignment as it directly estimates body mesh from accurate 3D keypoints.

The ground-truth shape annotations used by the methods for body mesh estimation [12, 27, 30] are usually obtained using the seminal approach MOSH [42]. MOSH [42] shows that a sparse set of 3D marker locations on the human body are sufficient to capture body shape and soft-tissue deformations. Motivated by this, in this work, we propose to harness the superior localization ability of recent keypoint

978-1-6654-2688-6/21/$31.00 ©2021 IEEE
DOI 10.1109/3DV53792.2021.00078

regressors [23, 43, 55, 68, 70] and reconstruct full human body mesh from the regressed 3D keypoint positions only. A method with this capability offers two main advantages: 1) As compared to the traditional regression based methods [27, 30], the mesh estimates will be more accurate and align better with visual clues since the 3D keypoints can be localized more accurately from images [23, 43, 55, 68, 70]. 2) It does not require images paired with ground-truth shape labels which are very hard to acquire.

Some existing works propose solutions in this direction but formulate the problem as an optimization framework where the parameters of a parametric body model (*e.g.*, SMPL [41]) are optimized to match the articulation of a regressed 3D pose [48, 79]. Some other methods instead use 2D positions, but can also be extended to 3D [6, 52, 65]. Optimization-based methods are, however, prone to local-minima and are time consuming, in particular, when a parametric mesh model with large number of vertices (*e.g.* SMPL has 6890 vertices) has to be optimized. In contrast, in this work, we propose an analytical solution using a set of straightforward geometrical operations that are not prone to local-minima and have negligible computational cost while showing better performance.

We enable analytical mesh articulation by learning a 3D keypoint regressor that provides 3D positions of a sufficient number (K=26) of keypoints to accurately capture orientation of most body parts. The main challenge to learn such a regressor is to have ground-truth 3D annotations as the existing datasets do not provide annotations for enough keypoints. We show that such a regressor can be learned using synthetic and/or weakly-labeled data. For this, we build on the recent progress in weakly-supervised 3D keypoint learning [23, 29, 61], and train our keypoint regressor using a combination of fully- and weakly-labeled data. The keypoints that do not have labeled 3D annotations are learned using weakly-labeled data in the form of unlabeled multi-view images along with a collection of images annotated with 2D positions only.

Given the estimated 3D keypoint positions, we then present KAMA, which is an analytical method for Keypoint Aware Mesh Articulation. It uses a set of straightforward geometrical operations to articulate a canonical mesh using the regressed 3D keypoint positions. While KAMA already achieves state-of-the-art results, we further show that the meshes can be refined further by using a simple first-order optimization that removes the discrepancies between the regressed keypoints and articulated mesh. As shown in Fig. 1, our approach offers accurate mesh fitting and significantly better alignment as compared to the traditional regression based state-of-the-art method [30]. We evaluate our proposed approach on the challenging 3DPW and Human3.6M datasets where it achieves state-of-the-art results.

2. Related Work

In the following, we discuss existing methods for 3D keypoint regression and body mesh recovery.

3D Keypoint Regression: These methods regress 3D keypoint positions from an RGB image [13, 36, 37, 54, 55, 55, 59, 67, 68, 71–73, 83, 84] or a 2D pose [8, 19, 21, 45, 50, 62] as input. Recently, this is achieved by training a deep neural network using ground-truth 3D pose annotations. Earlier methods regress 3D keypoints using holistic regression with a fully-connected output layer [36, 37, 59, 71, 72, 84]. More recent methods, however, adopt fully-convolutional networks to produce volumetric [43, 55, 68, 70] or heatmap-like [23, 82] representations. This enables better correspondence between input image and the output 3D pose representation, and therefore, leads to higher localization accuracy. Since the acquisition of ground-truth 3D data is very hard, many recent works try to learn 3D keypoint regressors in semi [38, 58, 60, 61, 78, 80] and weakly [9, 15, 23, 29, 33, 49, 51, 56, 76, 77] supervised ways. In this work, we build on the advances of these methods to learn additional 3D keypoints required for our method.

Body Mesh Recovery: These methods estimate body pose as well as its shape from RGB images. Most of the recent works adopt deep neural networks and directly regress the parameters of a parametric body model, SMPL [41], from images [12, 17, 25, 27, 28, 30, 32, 53, 57, 63, 69, 85]. However, learning this non-linear mapping is very hard and often results in meshes that do not align very well with image content. Some recent methods [10, 31, 39] try to alleviate this problem by directly predicting the vertex coordinates from image features. However, the main limitations of these methods is that they rely heavily on ground-truth body shape annotations which are very hard to acquire.

Other works try to decompose the problem into stages. They first estimate 2D and/or 3D keypoints from images and then estimate the mesh parameters using optimization based methods [4, 6, 26, 52, 65, 79] or use graph CNN to directly reconstruct the mesh [11]. These methods rely on large collection of motion capture data, *e.g.*, AMASS [44], to learn strong body pose prior. The optimization based methods are, however, prone to local-minima due to 2D-3D depth ambiguities and require careful initialization for optimal solutions. Also, they can be computationally very intensive due to their iterative nature, in particular, when a body mesh with a large number of vertices has to optimized. Our work also falls into this category in that we first estimate the 3D body keypoints and then reconstruct the body mesh. However, we propose an analytical solution to articulate a canonical mesh using the estimated 3D keypoints and a set of geometric operations. Our proposed approach is neither expensive, nor it is prone to local-minima. Similar to our method, the contemporaneous work HybrIK [35] also reconstructs 3D body meshes from 3D keypoints. However,

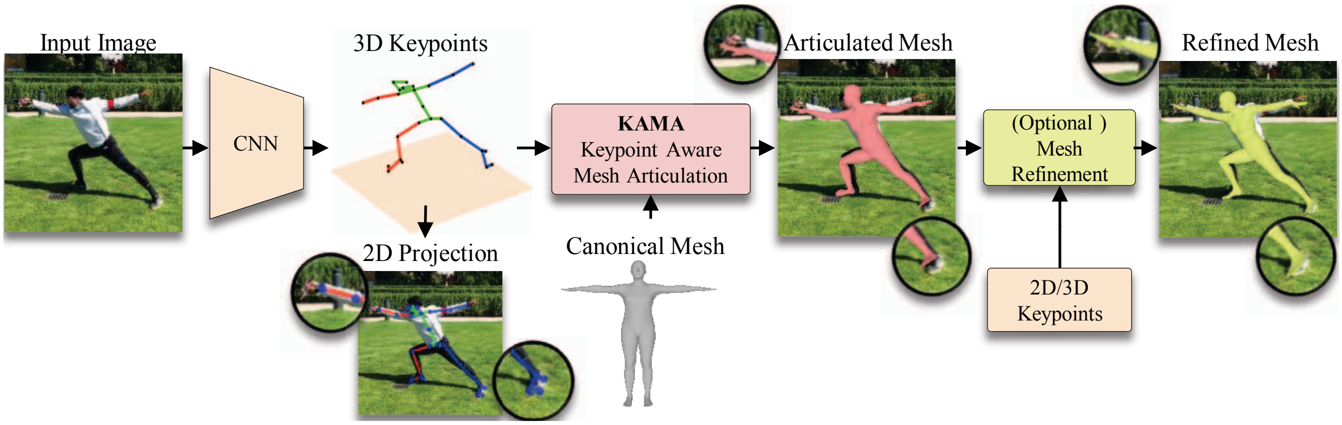

Figure 2. **Overview:** Given an RGB image as input, we use a 3D keypoint regressor that produces absolute 3D positions of 26 body keypoints. We use the estimated 3D positions to articulate the SMPL body model using a set of geometrical transformation (Sec. 3.2). While KAMA already provides very good mesh reconstruction, the estimated mesh can be (optionally) improved further using a very simple first order optimization that minimizes the discrepancies between the articulated mesh and the regressed 2D and 3D keypoints (Sec. 3.3). Our mesh estimates align much better with the image content as compared to the state-of-the-art methods.

in contrast to our method, it requires images annotated with 3D meshes.

Some methods adopt a hybrid approach by utilizing a regression followed by optimization strategy. The approaches [17, 81] train regressors that produce several pose representations in addition to the parameters of SMPL. These representations are then used in an optimization framework to refine the initial SMPL predictions. This strategy is, however, extremely data hungry. In addition to the labels for body pose and shape, they also require segmented part labels and DensePose [18] annotations. Our approach can also benefit from this hybrid-strategy as we will show in the experiments. However, in contrast to these methods, our approach does not require 3D mesh annotations or any kind of additional labels such as part segmentation or DensePose.

3. Method

Our goal is to reconstruct the full 3D body mesh $\mathbf{M}$ from an RGB image $\mathbf{I}$ of a pre-localized person. We do this via a set of 3D body keypoints that are obtained using a learned keypoint regressor. In the following, we first describe our approach for 3D keypoint regression (Sec. 3.1) and then present our proposed method for body mesh articulation from the regressed keypoints (Sec. 3.2). In (Sec. 3.3), we show that the estimated meshes can be refined further using a simple optimization objective and body mesh priors. An overview of the proposed approach can be seen in Fig. 2.

3.1. 3D Keypoint Regression

Our goal is to learn a keypoint regressor $\mathcal{F}(\mathbf{I})$, in the form of a deep neural network, that takes an image $\mathbf{I}$ as input and produces the 3D positions $\mathbf{X}=\{\mathbf{x}_k\}_{k \in K}$ of K body keypoints. Since we aim to use the keypoints to articulate a canonical mesh, the number of keypoints should be sufficient to obtain finer details about body mesh such as the head and feet orientation. The commonly used 17 keypoints are, however, insufficient for this purpose. For example, the 3D head pose (yaw, pitch, roll) in the mesh cannot be fully determined by the 3D positions of the neck and top-of-the-head locations only. We need additional 3D keypoints on the face to describe the full head pose. Therefore, in this work, we learn to regress $K{=}26$ body keypoints including, eyes, ears, nose, small and big toes, and heels, in addition to the other commonly used body keypoints. One main challenge to learn such a regressor is that the existing datasets for 3D human pose, such as the Human3.6M [20] and MPII-INF-3DHP [46], do not provide ground-truth annotations for the additional keypoints. To this end, we adopt the recent work of Iqbal *et al.* [23] that trains the 3D regressor using weakly-labeled data through multi-view consistency and 2D pose labels. During training, we supervise the keypoints that have ground-truth 3D annotations using fully-supervised losses and train the remaining 9 keypoints (eyes, nose, ears, toes, and heels) using using weakly-supervised losses via multi-view consistency as done in [23]. Such a joint fully and weakly-supervised training strategy allows us to train a 3D pose regressor with all $K{=}26$ keypoints given 2D annotations for all keypoints in one dataset (*i.e*., MS-COCO [7, 40]), and 3D annotations for some keypoints along with multi-view images from another dataset (*i.e*. Human3.6M [20]).

There are two additional useful properties of the keypoint regressor trained using [23]. First, it provides absolute 3D positions of the keypoints. Therefore, we recover the body mesh in the absolute camera space and use perspective-projection to project it onto the image plane. Second, it reconstructs 3D keypoints using a 2.5D heatmap representation [22], which yields 3D keypoints that are well-aligned with the image content. An example of our estimated 3D keypoints can be seen in Fig. 2. We refer the reader to [23] for further details about training the regressor.

3.2. KAMA: Keypoint Aware Mesh Articulation

Given the regressed 3D keypoints $\mathbf{X}$ from the previous section, our goal is to articulate a canonical mesh such that it matches the pose of the person. We encode the body mesh using the Skinned Multi-Person Linear (SMPL) model [41]. SMPL represents the body mesh using a linear function $M(\theta, \beta)$ that takes as input the pose parameters $\theta \in \mathbb{R}^{24\times 3}$ and the shape parameters $\beta \in \mathbb{R}^{10}$ and produces an articulated triangle mesh $\mathbf{M} \in \mathbb{R}^{V\times 3}$ with V=6980 vertices. The pose parameters θ consist of local 3D-rotation matrices, in axis-angle format, corresponding to each joint in the pre-defined kinematic structure ϵ of the human body. In the following, we estimate the pose parameters θ of the mesh from the regressed 3D keypoints using a set of geometrical transformations. We use a simple procedure that is fully analytic and the computational cost is completely negligible.

3.2.1 Keypoint Rotations from 3D Positions

Let $\bar{\mathbf{M}}$ be the body mesh in the canonical pose and $\bar{\mathbf{X}}{=}\mathbf{W}\bar{\mathbf{M}}{=}\{\bar{\mathbf{x}}_k\}_{k\in K}$ be the 3D keypoint positions in the canonical pose. Here $\mathbf{W} \in \mathbb{R}^{K\times V}$ is a learned weight matrix that defines the contribution of every vertex to the keypoints. Our goal is to use $\mathbf{X}$ and $\bar{\mathbf{X}}$ to calculate a set of rotations $\hat{\theta}{=}\{\theta_k\}_{k\in K}$ such that the mesh $\hat{\mathbf{M}}{=}M(\xi(\hat{\theta}), \beta{=}\mathbf{0}^{1\times 10})$ has an articulation similar to that of the regressed keypoints $\mathbf{X}$. Here the function $\xi(.)$ converts the order of rotation matrices from the 26-keypoint skeleton structure used for the keypoint regressor to the skeleton of SMPL which has 24 keypoints. Following the definition of SMPL, we use axis-angle representation of the rotation matrices. We define $C(k)$ as the children keypoints of keypoint k and $N(k)$ as the set of all keypoints adjacent to k, including k, as defined by the kinematic structure ϵ.

We apply three different rules to compute an initial estimation of the global rotations θ^g_k for every keypoint k: 1) For keypoints with one child we estimate rotation with ambiguous twist which is later compensated, 2) for keypoints with multiple children we estimate rotation with the help of other connected joints that move rigidly with k, and 3) we assume no rotation for childless keypoints. These rules are summarized as follows:

$$\theta^g_k = \begin{cases} \alpha_1(\bar{\mathbf{x}}_{c(k)}{-}\bar{\mathbf{x}}_k, \mathbf{x}_{c(k)}{-}\mathbf{x}_k) & \text{if } |C(k)| = 1 \\ \alpha_2(\bar{\mathbf{X}}^{\mathrm{N}}_k, \mathbf{X}^{\mathrm{N}}_k) & \text{if } |C(k)| > 1 \\ \mathbf{0}^{1\times 3} & \text{otherwise,} \end{cases} \quad (1)$$

where $\bar{\mathbf{X}}^{\mathrm{N}}_k = \{\bar{\mathbf{x}}_n\}_{n\in N(k)}$ and $\mathbf{X}^{\mathrm{N}}_k = \{\mathbf{x}_n\}_{n\in N(k)}$.

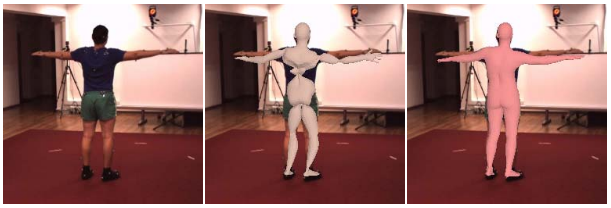

Figure 3. Illustration of ambiguous rotations for the keypoints with one child. Left: Input image. Middle: The articulated mesh has ambiguous twists around all keypoints with one child. Right: The articulated mesh after the removal of ambiguous twists. Both meshes are articulated using the same 3D keypoints.

Keypoints with one child: For the keypoints with one child, we compute rotation as the angle applied to the vector perpendicular to the plane formed by the bones $\bar{\mathbf{x}}_{c(k)}{-}\bar{x}_k$ and $\mathbf{x}_{c(k)}{-}\mathbf{x}_k$ in the canonical and estimated poses, respectively. $c(k)$ corresponds to the index of the child of keypoint k, and the function $\alpha_1(\mathbf{v}_1, \mathbf{v}_2)$ provides the rotation in axis-angle format as follows:

$$\alpha_1(\mathbf{v}_1, \mathbf{v}_2) = \arccos\Big(\frac{\mathbf{v}_1^T\mathbf{v}_2}{||\mathbf{v}_1||\,||\mathbf{v}_2||}\Big) \cdot \frac{\mathbf{v}_1 \times \mathbf{v}_2}{||\mathbf{v}_1 \times \mathbf{v}_2||}, \quad (2)$$

where the right part represents the axis of rotation and the left part corresponds to the angle of rotation.

It is important to note that the rotation estimated in this way is inherently ambiguous as any arbitrary twist about the child vector can be applied without affecting the position of the child keypoint. We will remove such ambiguous twists after calculating rotations for all keypoints as explained later in this section.

Keypoints with multiple children: For the keypoints with multiple children, we can estimate the keypoint rotation more precisely. Here we assume that all keypoints in $N(k)$ move rigidly, and estimate the rotation as a rigid rotation between $\bar{\mathbf{X}}^{\mathrm{N}}_k$ and $\mathbf{X}^{\mathrm{N}}_k$ as

$$\alpha_2(\bar{\mathbf{X}}^{\mathrm{N}}_k, \mathbf{X}^{\mathrm{N}}_k) = \underset{\theta}{\operatorname{argmin}} \sum_{\substack{\bar{\mathbf{x}}_i \in \bar{\mathbf{X}}^{\mathrm{N}}_k \\ \mathbf{x}_i \in \mathbf{X}^{\mathrm{N}}_k}} \psi(\mathbf{x}_i)(\phi(\theta, \bar{\mathbf{x}}_i) - \mathbf{x}_i), \quad (3)$$

where $\phi(\theta, \bar{\mathbf{x}})$ represents rotating the vector $\mathbf{x}$ with θ using Rodriguez formula, and $\psi(\mathbf{x_i})$ corresponds to the detection confidence of keypoint i as provided by the keypoint regressor $\mathcal{F}(\mathbf{I})$. The eq. (3) can be solved easily in closed-form using singular value decomposition [16]. Thanks to our 26-keypoint regressor, many of the keypoints (*i.e.*, pelvis, neck, nose/face, ankles) fall into this category.

Global to local rotations: The rotations of the body joints as calculated above are the global rotations for each of them. However, to be able to use them in the function $M(\xi(\theta), \beta)$ to articulate the SMPL mesh, we convert them to local rotations as follows:

$$\theta_k = \theta^{g-1}_{p(k)} \cdot \theta^g_k, \quad (4)$$

where $p(k)$ is the index of the parent of keypoint k. The root keypoint has no parent so it remains unchanged.

Twist removal: Given the local rotations for all body joints, we need to remove unnecessary twists from the rotations of the joints with one child. A reasonable choice is to default to the twist from the canonical pose (which is zero by definition). This can be done via swing-after-twist decomposition [14]. Specifically, we decompose the estimated local rotation into its swing and twist components, and then set the rotation as the swing component, while discarding the twist component. A comparison between the meshes before and after the twist removals can be seen in Fig 3.

3.2.2 Scale and Translation Estimation

So far, we have articulated the canonical mesh to match the pose of our regressed keypoints. However, it still lies at the origin and its global scale is unknown. Since our keypoint regressor provides absolute 3D pose including approximate bone length scales, we calculate the global translation $\mathbf{t} \in \mathbb{R}^3$ and global scale $s \in \mathbb{R}$ for the articulated mesh using Procrustes analysis [16] between the keypoints of the mesh and regressed keypoints:

$$\hat{s}, \hat{\mathbf{t}} = \underset{s,\mathbf{t}}{\operatorname{argmin}} ||\mathbf{W}(s\hat{\mathbf{M}}+\mathbf{t}) - \mathbf{X}||_2^2. \quad (5)$$

where $\hat{\mathbf{M}}{=}M(\xi(\hat{\theta}), \beta{=}\mathbf{0}^{1\times 10})$ is the articulated mesh using the estimated rotations. This gives us our final articulated mesh in the absolute camera coordinate system as $\mathbf{M}{=}\hat{s}\hat{\mathbf{M}}{+}\hat{\mathbf{t}}$. We use perspective-projections to project the resulting mesh onto the image plane.

3.3. Pose Refinement and Shape Estimation

While our approach for mesh articulation using keypoints already provides state-of-the-art mesh estimates, as we will show later in our experiments, there are a few issues that can be addressed further. First, our method resorts to canonical twist for the keypoints with single children which is not the most optimal choice. Second, the regressed keypoints do not exactly match with the skeleton structure of SMPL. For example, in contrast to SMPL, the regressor does not provide any keypoints on the collar bones. Depending on the 2D annotations, there can be other subtle differences between the estimated keypoints and the keypoints in the canonical mesh. Also, small errors in one keypoint can propagate to the entire mesh. For example, an incorrect rotation for pelvis will impact all other keypoints and will result in mesh keypoints that are very different from the regressed keypoints. Lastly, we also need to estimate the shape parameters β of SMPL to fully capture the body details. To this end, we build on [6] and remove such discrepancies by using body pose and shape priors in an energy minimization formulation that further refines the pose parameters θ, shape β, global translation $\mathbf{t}$, and global scale s:

$$\hat{\theta}, \hat{\beta}, \hat{\mathbf{t}}, \hat{s} = \underset{\theta,\beta,\mathbf{t},s}{\operatorname{argmin}} \mathcal{L}(\theta, \beta, \mathbf{t}, s), \quad (6)$$

where $\mathcal{L}(\theta, \beta, \mathbf{t}, s)$ consists of four errors terms

$$\mathcal{L}(\theta, \beta, \mathbf{t}, s) = \mathcal{L}_{\text{2D}} + \omega_1 \mathcal{L}_{\text{3D}} + \omega_3 \mathcal{L}_\theta + \omega_2 \mathcal{L}_\beta. \quad (7)$$

The error term $\mathcal{L}_{\text{2D}}$ is the reprojection error. It measures the discrepancies between the 2D keypoints provided by the regressor and the projection of the mesh skeleton $\hat{\mathbf{X}}{=}\{\hat{\mathbf{x}}_k\}_{k\in K}{=}\mathbf{W}(s\mathbf{M}+\mathbf{t})$:

$$\mathcal{L}_{\text{2D}} = \sum_k \psi(\mathbf{x}_k) ||P(\mathbf{K}, \mathbf{x}_k) - P(\mathbf{K}, \hat{\mathbf{x}}_k)||_2^2, \quad (8)$$

where $\mathbf{K}$ is the intrinsic camera matrix, $P(.,.)$ represents projection on the image plane, $\psi(\mathbf{x}_k)$ corresponds to the detection score of the keypoint k. $\mathcal{L}_{\text{3D}}$ measures the difference between predicted 3D position and the 3D mesh skeleton:

$$\mathcal{L}_{\text{3D}} = \sum_k \psi(\mathbf{x}_k) ||\mathbf{x}_k - \hat{\mathbf{x}}_k||_2^2. \quad (9)$$

The error terms $\mathcal{L}_\theta$ and L_β correspond to the pose prior and shape prior terms as defined in [6], respectively. Specifically, $\mathcal{L}_\theta$ favors plausible pose parameters θ. In our case, it helps in recovering the optimal twist for keypoints with one child and in reducing the ambiguities due to missing keypoints and differences in the skeleton structures. The term $\mathcal{L}_\beta$ is a regularization for parameters β such that the optimized shape is not distant from the mean shape.

For the optimization, we use the values of θ, s, and $\mathbf{t}$ from KAMA as initialization and use Adam as the optimizer. Since we start from a very good initialization, we found that the optimization converges within 100 iterations without the need of a multi-stage optimization strategy as required by prior works [6,52]. Some examples of mesh estimates before and after the refinement can be seen in Fig. 4

4. Implementation Details

We follow [23] and use HRNet-w32 [66] as the keypoint regressor. We empirically choose $\omega_1 = 500$, and adopt $\omega_2 = 4.78$ and $\omega_3 = 5$ from [6, 30]. We use the publicly available implementation of SMPL provided by [30]. The linear regressor $\mathbf{W}$ in this implementation allows to extract 54 keypoints from the mesh vertices. We choose 26 keypoints that are closest to the 2D annotations used by our keypoint regressor. Note that these keypoints do not exactly overlap with the native 24 keypoints of SMPL, but are sufficient to calculate enough rotation matrices in $\theta \in \mathbb{R}^{24\times 3}$ to capture the full body pose. The rotation matrices that cannot be estimated (*i.e.*, collar-bones, spine-1, spine-3, and hands) are assigned zeros in (1), but optimized in (6).

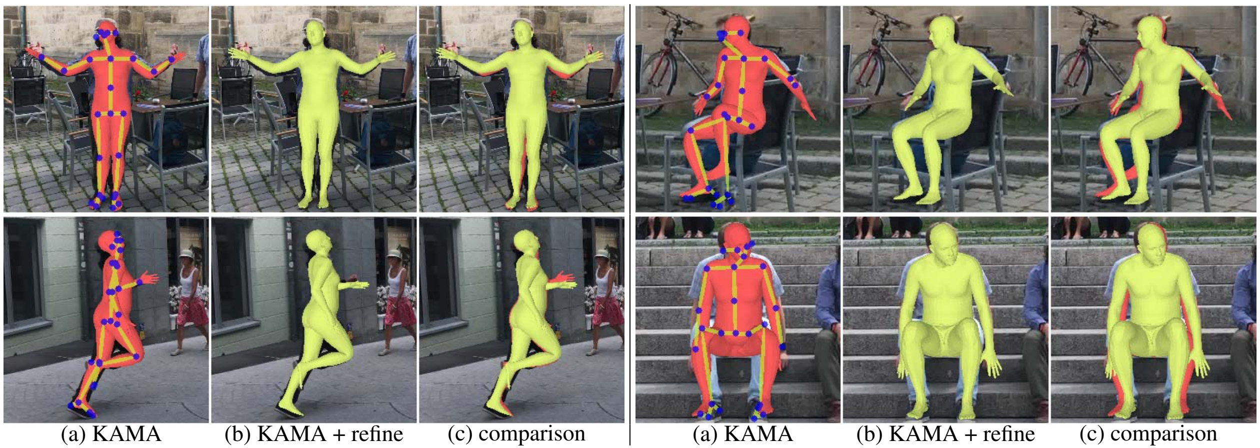

Figure 4. (a) Articulated body meshes obtained using our proposed approach KAMA (Sec. 3.2). (b) Meshes obtained after pose and shape refinement (Sec. 3.3). (c) Comparison between a & c. While KAMA already provides very good mesh estimates, they sometimes can have errors due to missing twist information, errors in the regressed keypoints, error propagation, occlusions, and etc. Such errors can be fixed using a simple optimization objective consisting of body pose and shape priors.

5. Experiments

We evaluate the performance of the proposed approach in detail and also compare it with the state-of-the-art image-based methods for human pose and shape estimation.

5.1. Datasets

Human3.6M [20]: We follow the standard protocol [27] and use five subjects (S1, S5, S6, S7, S8) for training and test on two subjects (S9 and S11) on the frontal camera.

3D Poses in-the-Wild (3DPW) [75]: consists of 60 videos recorded in diverse environments. We follow the standard protocol and use its test-set for evaluation and do not train on this dataset.

RenderPeople: This is a synthetic dataset with ground-truth annotations for all K=26 keypoints used in our method. We used 10 characters from RenderPeople [2] dataset and generated 80k images under a variety of poses using CMU MoCap dataset [3] while using ∼100 outdoor HDRI image from HDRI Haven [1] for lighting and backgrounds. We manually annotated the vertices corresponding to eyes, ears, nose, toes and heels for each character as they are not part of the rigged skeletons.

MS-COCO (COCO) [40]: We use this dataset as the weakly-labeled set for the training of the 3D keypoint regressor. The dataset provides 2D annotations for 18 keypoints. A subset of the dataset was augmented by [7] with annotations for 3 additional keypoints on each foot.

5.2. Evaluation and Training Setting

We report Mean Per-Vertex Error (MPVE) and 3D reconstruction error in millimeters (mm) for all experiments. Following the standard practice [27], we extract 14 keypoints for evaluation from the recovered mesh using a pre-trained linear regressor.

For 3DPW dataset, different methods use different datasets for training which include Human3.6M [20], MSCOCO [40], MPI-INF-3DHP [46], MuCo-3DHP [47] MPII [5], LSP [24], UP [34], SURREAL [74] and etc. In this work, we only use Human3.6M, MSCOCO and the 80k synthetic images from RenderPeople dataset to train the model used for evaluation on 3DPW dataset.

For evaluation on Human3.6M, we train only using Human3.6M and MSCOCO datasets. The ground-truth mesh annotations for Human3.6M are only available to a sparse set of researchers as the distribution has been discontinued. Hence, we only report reconstruction error for Human3.6M. We also found that there are discrepancies between the 14 keypoints extracted using the keypoint regressor provided by [6] and the ground-truth marker locations from Human3.6M. This is not a problem when ground-truth mesh annotations are available as the consistent ground-truth keypoints can be extracted from the meshes. Since the mesh annotations are not available to us, we remove these discrepancies by training another linear regressor (42×42 weight matrix) using training data, and apply it to the extracted 14 keypoints before evaluation.

5.3. Ablation Study

In Tab. 1, we evaluate all components of the proposed approach. We chose 3DPW datasets for all ablative studies as it represents more general in-the-wild scenarios, and also provides ground-truth mesh annotations. First, we evaluate the performance of our approach for mesh articulation (KAMA) using regressed 3D keypoints. If we do not remove ambiguous twists from the calculated rotations, as explained in Sec 3.2.1, the estimated meshes yield a MPVE and 3D reconstruction error of 124.8mm and 64.0mm, respectively. Removing the ambiguous twists results in a significant decrease in the error (124.8mm vs. 107.7mm

Methods	**MPVE**	**Recon. Error**
articulation using 3D keypoints (1)		
KAMA w/o twist removal	124.8	64.0
KAMA with twist removal	107.7	54.5
pose & shape refinement using (6)		
Initialization using (1)		
$\mathcal{L}_{2D}$	152.5	87.8
$\mathcal{L}_{2D}$ + $\mathcal{L}_{3D}$	115.8	63.5
$\mathcal{L}_{2D}$ + $\mathcal{L}_{3D}$ + $\mathcal{L}_{\theta}$	100.4	53.0
$\mathcal{L}_{2D}$ + $\mathcal{L}_{3D}$ + $\mathcal{L}_{\theta}$ + $\mathcal{L}_{\beta}$	97.0	51.1
No initialization	106.6	56.4
Init. using mean pose (SMPLify3D)	100.7	55.8
Init. using SPIN [30]	94.8	50.4
Init. using mean pose - w/o $\mathcal{L}_{2D}$ (SMPLify2D)	115.0	69.3
Init. using SPIN [30] - w/o $\mathcal{L}_{2D}$	98.8	55.0
impact of 3D keypoint quality - using GT 3D keypoints		
KAMA	47.4	18.0
KAMA with refinement using (6)	44.1	17.0

Table 1. Impact of different components in the proposed approach.

and 64.0mm vs. 54.5mm) and shows the importance of this step. Note that ambiguous twists have higher impact on the MPVE since the body surface is impacted more with incorrect twist rotations as compared to body keypoint positions. We would like to emphasize that the errors of 107.7mm and 54.5mm are the state-of-the-art on 3DPW dataset, even though we only use mean shape values (*i.e.*, $\beta = \mathbf{0}^{1\times10}$)) in KAMA. Thanks to eq. (5), we can find an optimal global scale for the mesh without having to optimize beta.

Next, we evaluate the contributions of different error terms used in (7). In all cases we initialize the body joint rotations using (1) and body translation and scale using (5). If we only optimize for the re-projection loss $\mathcal{L}_{2D}$, the errors increase significantly (from 107.7mm to 152.5mm and 54.5mm to 87.8mm) due to the well known 2D-3D ambiguities. Adding $\mathcal{L}_{3D}$ reduces the errors (from 152.2 to 115.8mm and 87.8mm to 63.5mm) but remains higher than what we can achieve by using KAMA only. This is because optimizing 2D and 3D losses only is still susceptible to ambiguous twists. In KAMA, on the other hand, we explicitly handle the twist by either discarding it or estimating it with the help of adjacent keypoints. Enforcing body pose priors using $\mathcal{L}_{\theta}$ significantly reduces the errors. As compared to KAMA only, the MPVE and joint reconstruction errors are reduced from 107.7mm to 100.4mm and 54.5mm to 53.0mm, respectively. As mentioned earlier, $\mathcal{L}_{\theta}$ encourages plausible poses. In our case, it helps to recover the twists of keypoints with single child and to reduce the ambiguities due to missing 3D keypoints and differences in the skeleton structures. Finally, adding $\mathcal{L}_{\beta}$ results in further decrease in the errors demonstrating the importance of optimal body shape parameters.

To emphasize the usefulness of KAMA for the optimization based method, we also evaluate the case when no initialization for θ is used. Since we have the estimated 3D keypoints, we can obtain reasonable initial values for global scale s, global translation $\mathbf{t}$ and the global orientation θ_0 by calculating a rigid transformation between the regressed keypoints and skeleton of the canonical mesh using Procrustes analysis. We initialize β and θ with zeros. Optimizing (6) without a good initialization results in a 3D error of 56.4mm which is significantly higher than the case when KAMA is used as the initialization (51.1mm), demonstrating the importance of KAMA and accurate initialization.

Methods		Mesh Supervision	MPVE	Recon. Error
SMPLify [6]*	ECCV'16	N	-	106.1
HMR	CVPR'18	Y	161.0	81.3
Kundu *et al.* [32]	ECCV'20	N	-	78.2
ExPose [12]	ECCV'20	Y	-	60.7
Rong *et al.* [63]	CVPR'19	Y	152.9	-
SPIN [30]	ICCV'19	Y	112.8	59.2
Pose2Mesh [11]	ECCV'20	Y	-	58.9
I2L-Mesh [10]	ECCV'20	Y	110.1	58.6
Zanfir *et al.* [81]	ECCV'20	Y	-	57.1
Song *et al.* [65]*	ECCV'20	N	-	55.0
KAMA (ours)		N	**107.7**	**54.5**
KAMA w. refinement*		N	**97.0**	**51.1**

Table 2. Comparison with state-of-the-art methods on **3DPW** dataset. *optimization-based methods

To further evaluate the impact of initialization, we implement a 3D version of SMPLify [6] which initializes the pose parameters θ with the mean body pose. We initialize the global orientation using the rotation of the pelvis keypoint obtained using (1) and global scale and translation using (5). This results in a MPVE and 3D reconstruction error of 100.7mm and 55.8mm, respectively, that are significantly higher than the case when predictions from KAMA are used for initialization, demonstrating that KAMA serves as a very good initialization. In fact, KAMA without any optimization achieves better 3D reconstruction error than SMPLify3D (54.5 vs 55.8mm). We also evaluate when an off-the-shelf regression based method, SPIN [30], is used as initialization. Even though SPIN uses full 3D pose and shape supervision, using its predictions as initialization achieve results on par with KAMA. We also report a 2D version of optimization by removing L_{3D} from the objective while keeping all other error terms and initialization as before. The errors increase significantly showing that the 3D keypoints are important for accurate reconstruction. In contrast to SPIN [30], KAMA by default exploits the strengths of 3D keypoints. It significantly outperforms SPIN when no refinement is performed (59.2 vs 54.5mm, see Tab. 2).

Finally, we also evaluate the impact of 3D keypoint accuracy on body mesh reconstruction using KAMA. For this, we extracted 26 keypoints from the ground-truth meshes, and used KAMA to reconstruct body mesh. This setting serves as an upper-bound for KAMA. Using ground-truth 3D keypoints significantly decreases the errors showing that the performance can be improved further by using a more accurate 3D keypoint regressor. Notably, the dif-

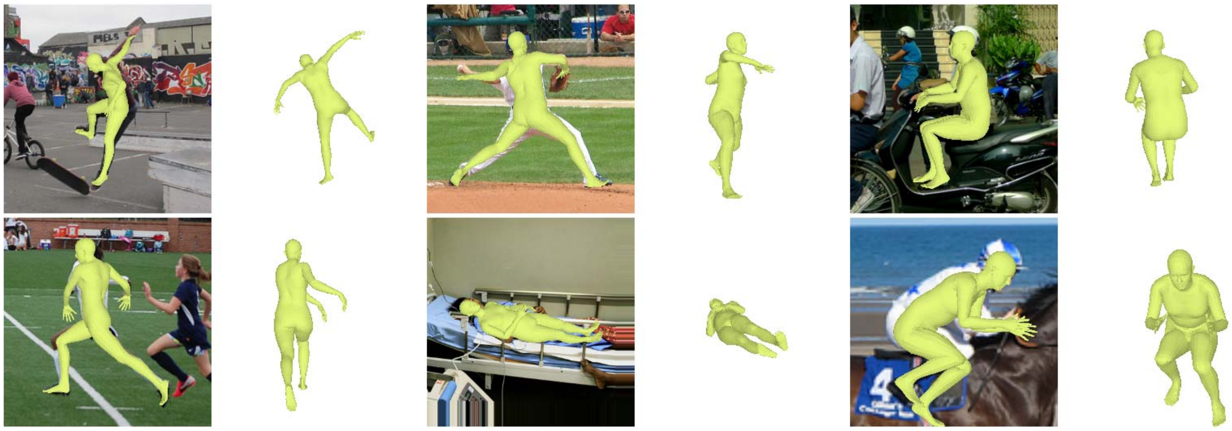

Figure 5. Some qualitative results from the validation set of COCO dataset.

ference between KAMA and KAMA-with-refinement decreases which indicates that the refinement step may not be required with more accurate 3D keypoints.

5.4. Comparison to the State-of-the-Art

We compare the performance of our approach with the state-of-the-art on 3DPW and Human3.6M datasets.

Tab. 2 compares our proposed method with the state-of-the-art on 3DPW dataset. We chose the best numbers reported in all papers. The MPVE for SPIN [30] and I2L-Mesh [10] are obtained using the publicly available source codes, whereas the MPVE for HMR is obtained from [63]. KAMA without any additional refinement already outperforms all state-of-the-art methods. Refining the mesh estimates using (6) further improves the results and sets a new state-of-the-art on 3DPW dataset. Note that the methods [10, 17, 27, 30, 31, 53, 81] use images annotated with ground-truth mesh annotations, and the method [65] relies on very strong pose priors learned from a massive corpus of MOSHed [42] motion capture data, AMASS [44]. KAMA, in contrast, does not require any mesh annotations and is trained using 3D keypoints supervision only, yet it outperforms them with a large margin. This demonstrates that a good keypoint regressor combined with KAMA can yield state-of-the-art mesh reconstruction without the need of hard-to-acquire body mesh annotations.

Tab. 3 compares our proposed method with the state-of-the-art on Human3.6M dataset. On this dataset, KAMA without refinement performs on-par with SPIN [30] and I2L-MeshNet [10]. This is likely because of the limited diversity of Human3.6M where these methods can overfit using the full mesh annotations. Nonetheless, as before, refining the predictions of KAMA using (6) reduces the errors and results in state-of-the-art performance on Human3.6M. We would also like to emphasize that KAMA, unlike many other methods, provides meshes in absolute camera coordinates and uses perspective projection to project the meshes on to the images. This is in contrast to *e.g.*, SPIN [30] and other methods that only predict root-relative meshes and use weak-perspective projection, hence, incur lower errors as compared to our method.

Methods		Mesh Supervision	Reconstruction uError
SPMLify [6]*	ECCV'16	N	82.3
SMPLify-X [52]*	CVPR'19	N	75.9
HMR [27]	CVPR'18	Y	56.8
Song *et al*. [65]*	ECCV'20	N	56.4
GraphCMR [31]	CVPR'19	Y	50.1
STRAPS [64]	BMVC'20	N	55.4
Pose2Mesh [11]	ECCV'20	Y	47.0
TexturePose [53]	ICCV'19	Y	49.7
Kundu *et al*. [32]	ECCV'20	N	48.1
HoloPose [17]	CVPR'19	Y	46.5
DSD [69]	ICCV'19	Y	44.3
I2L-MeshNet [10]	ECCV'20	Y	41.7
SPIN *et al*. [30]	ICCV'19	Y	41.1
Ours		N	**41.5**
Ours w. refinement*		N	**40.2**

Table 3. Comparison with the state-of-the-art methods on **Human3.6M** dataset. *optimization-based methods

Finally, in Fig. 5, we provide some qualitative results of our approach on in-the-wild images.[1]

6. Conclusion

In this work, we presented a novel approach for human mesh recovery from 3D keypoint only. To this end, we used a 3D keypoint regressor that is able to estimate 3D positions of 26 body keypoints. We then presented, KAMA, a 3D keypoint aware approach to articulate a canonical mesh using 3D keypoint positions and a set of simple geometrical operations. We then further improved the mesh estimates via a pose refinement and shape estimation approach. The resulting meshes are accurate and align well with image content. In contrast to existing methods, our approach does not require 3D body shape annotations and provides meshes in the absolute camera coordinates. Yet, it achieves state-of-the-art results on the challenging benchmarks.

[1] More qualitative results: https://youtu.be/mPikZEIpUE0

References

[1] HDRI Haven, 2020. https://renderpeople.com/3d-people. 6

[2] Render People, 2020. https://hdrihaven.com/. 6

[3] Carnegie mellon university graphics lab: Motion capture database, 2014. 6

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

[5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 6

[6] 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 *ECCV*, 2016. 2, 5, 6, 7, 8

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2019. 3, 6

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

[9] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *CVPR*, 2019. 2

[10] Hongsuk Choi, 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 *ECCV*, 2020. 1, 2, 7, 8

[11] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 2, 7, 8

[12] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 1, 2, 7

[13] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *ECCV*, 2018. 2

[14] P. Dobrowolski. Swing-twist decomposition in clifford algebra. *ArXiv*, abs/1506.05481, 2015. 5

[15] Dylan Drover, Rohith M. V, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *ECCV Workshops*, 2018. 2

[16] David Wayne Eggert, A Lorusso, and Robert Bob Fisher. Estimating 3-d rigid body transformations: A comparison of four major algorithms. In *Machine Vision and Applications*, 1997. 4, 5

[17] Rıza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, 2019. 1, 2, 3, 8

[18] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *CVPR*, 2018. 3

[19] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3d pose estimation. In *ECCV*, 2018. 2

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 3, 6

[21] Umar Iqbal, Andreas Doering, Hashim Yasin, Björn Krüger, Andreas Weber, and Juergen Gall. A dual-source approach for 3D human pose estimation in single images. *CVIU*, 2018. 2

[22] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via 2.5D latent heatmap regression. In *ECCV*, 2018. 3

[23] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *CVPR*, 2020. 1, 2, 3, 4, 5

[24] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 6

[25] 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. 2

[26] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 2

[27] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 6, 8

[28] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2

[29] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *CVPR*, 2019. 2

[30] 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 *ICCV*, 2019. 1, 2, 5, 7, 8

[31] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 2, 8

[32] Jogendra Nath Kundu, Mugalodi Rakesh, Varun Jampani, Rahul Mysore Venkatesh, and R. Venkatesh Babu1. Appearance consensus driven self-supervised human mesh recovery. In *ECCV*, 2020. 2, 7, 8

[33] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *CVPR*, 2020. 2

[34] 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 *CVPR*, 2017. 6

[35] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021. 2

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

[37] Sijin Li, Weichen Zhang, and Antoni Chan. Maximum-margin structured learning with deep networks for 3D human pose estimation. In *ICCV*, 2015. 2

[38] Zhi Li, Xuan Wang, Fei Wang, and Peilin Jiang. On boosting single-frame 3d human pose estimation via monocular videos. In *ICCV*, 2019. 2

[39] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2

[40] 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 *ECCV*, 2014. 3, 6

[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 34(6):248:1–248:16, 2015. 2, 4

[42] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *SIGGRAPH Asia*, 33(6):220:1–220:13, 2014. 1, 8

[43] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018. 1, 2

[44] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 2, 8

[45] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 2

[46] 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 *3DV*, 2017. 3, 6

[47] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d body pose estimation from monocular rgb input. In *3DV*, 2018. 6

[48] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. In *SIGGRAPH*, 2017. 2

[49] Rahul Mitra, Nitesh B. Gundavarapu, Abhishek Sharma, and Arjun Jain. Multiview-consistent semi-supervised learning for 3d human pose estimation. In *CVPR*, 2020. 2

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

[51] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *ICCV*, 2019. 2

[52] 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 *CVPR*, 2019. 2, 5, 8

[53] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 2, 8

[54] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *CVPR*, 2018. 2

[55] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017. 1, 2

[56] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *CVPR*, 2017. 2

[57] 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*, 2018. 2

[58] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019. 2

[59] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2D and 3D human sensing. In *CVPR*, 2017. 1, 2

[60] Helge Rhodin, Mathieu Salzmann, , and Pascal Fua. Unsupervised geometry-aware representation learning for 3d human pose estimation. In *ECCV*, 2018. 2

[61] 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 *CVPR*, 2018. 2

[62] Matteo Ruggero Ronchi, Oisin Mac Aodha, Robert Eng, and Pietro Perona. It's all relative: Monocular 3d human pose estimation from weakly supervised data. In *BMVC*, 2018. 2

[63] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *ICCV*, 2019. 2, 7, 8

[64] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *BMVC*, 2020. 8

[65] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 2, 7, 8

[66] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 5

[67] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, 2017. 1, 2

[68] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 1, 2

[69] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, , and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, 2019. 2, 8

[70] István Sárándi, Timm Linder, Kai O. Arras, and Bastian Leibe. Metric-scale truncation-robust heatmaps for 3d human pose estimation. In *FG*, 2020. 1, 2

[71] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3D human pose with deep neural networks. In *BMVC*, 2016. 1, 2

[72] Bugra Tekin, Pablo Marquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2D and 3D image cues for monocular body pose estimation. In *ICCV*, 2017. 2

[73] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *CVPR*, 2017. 2

[74] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 6

[75] 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 *ECCV*, 2018. 6

[76] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *CVPR*, 2019. 2

[77] Chaoyang Wang, Chen Kong, and Simon Lucey. Distill knowledge from nrsfm for weakly supervised 3d pose learning. In *ICCV*, 2019. 2

[78] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. In *ECCV*, 2016. 2

[79] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body and hands in the wild. In *CVPR*, 2019. 2

[80] Yuan Yao, Yasamin Jafarian, and Hyun Soo Park. Monet: Multiview semi-supervised keypoint detection via epipolar divergence. In *ICCV*, 2019. 2

[81] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *ECCV*, 2020. 3, 7, 8

[82] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lv. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *ICCV*, 2019. 2

[83] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 2

[84] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *ECCV Workshops*, 2016. 1, 2

[85] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular realtime full body capture with inter-part correlations. In *CVPR*, 2021. 2