This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Exploiting temporal context for 3D human pose estimation in the wild

Anurag Arnab[1*†] aarnab@robots.ox.ac.uk
Carl Doersch[2*] doersch@google.com
Andrew Zisserman[1,2] zisserman@google.com

[1]University of Oxford [2]DeepMind

Abstract

We present a bundle-adjustment-based algorithm for recovering accurate 3D human pose and meshes from monocular videos. Unlike previous algorithms which operate on single frames, we show that reconstructing a person over an entire sequence gives extra constraints that can resolve ambiguities. This is because videos often give multiple views of a person, yet the overall body shape does not change and 3D positions vary slowly. Our method improves not only on standard mocap-based datasets like Human 3.6M – where we show quantitative improvements – but also on challenging in-the-wild datasets such as Kinetics. Building upon our algorithm, we present a new dataset of more than 3 million frames of YouTube videos from Kinetics with automatically generated 3D poses and meshes. We show that retraining a single-frame 3D pose estimator on this data improves accuracy on both real-world and mocap data by evaluating on the 3DPW and HumanEVA datasets.

1. Introduction

Understanding the 3D configuration of the human body has numerous real-life applications in robotics, augmented and virtual reality, and animation, among other fields. However, it is an inherently under-constrained problem when only a single image is available, as there are many 3D poses which project to the same 2D image. Data-driven methods to resolve this ambiguity are promising, but they are typically trained and evaluated on motion capture datasets recorded in constrained and unrealistic environments [17, 43, 29, 19].

To resolve some of the ambiguities in monocular 3D pose estimation, we exploit temporal consistency across frames of a video. The temporal dimension of ordinary videos encodes valuable information: multiple views of people are observed, where the body shape and bone lengths remain constant throughout a video, and joint positions in both 2D and 3D change slowly over time. These priors constrain the space of possible poses and thus help reduce the ambiguity of this ill-posed problem as shown in Fig. 1. Despite its value, the temporal information in mocap datasets is discarded by all current leading 3D pose estimation algorithms [20, 36, 45, 28] which use only single, ambiguous frames. Our approach incorporates temporal information through a form of *bundle adjustment*, a method used in multi-view geometry for estimating cameras and 3D structure of rigid scenes from image correspondences [13, 48]. We repurpose bundle adjustment to deal with non-rigid (articulated) human motion in a video sequence. In contrast to previous recurrent models for human pose [15], our method can jointly reason about all frames in the video, and errors made in initial frames do not accumulate over time. As illustrated in Fig. 1, the current state-of-art single frame estimation network for the SMPL model [20] fails on a number of frames of "in the wild" videos, such as when there is occlusion, unusual poses, poor lighting or motion blur. Our bundle adjustment method is able to correct these estimates and infer 3D human pose for these frames.

To address the lack of real-world data in 3D pose estimation, we apply our bundle adjustment framework to "in the wild" clips from the Kinetics dataset [22] comprised of YouTube videos, and show how we can leverage our predictions on real-world videos as a source of weak supervision to improve existing 3D pose estimation models. By encouraging temporal consistency with bundle adjustment and using YouTube videos as a source of weakly supervised data, we make the following novel contributions:

First, we show that multi-frame bundle adjustment can be specialized to human pose estimation, which improves performance on the Human 3.6M dataset over single frame estimation. Our method achieves the state-of-the-art for SMPL [26] models on this dataset.

We then apply our bundle adjustment method to 107 000 YouTube videos from the Kinetics dataset [22] and generate a large-scale dataset of 3D human poses aligned with the video frames. This dataset contains great diversity in pose, with 400 different human actions, and will be publicly released as a training resource via the DeepMind website. As

*Equal contribution.
†Work done during an internship at DeepMind

Figure 1. Although monocular 3D pose estimation is an ill-posed problem, state-of-art methods [20] do not use temporal information to constrain the problem. Coupled with the fact that 3D supervision is only available from lab-captured mocap datasets, they often fail on "in the wild" videos, e.g., from Kinetics [22]. As shown in the second row, the failure modes of [20] vary even though the image has barely changed. Our proposed bundle adjustment considers all frames in the video jointly and uses temporal coherence to prevent major failures (column 2 and 3) and to resolve ambiguities (column 5). We then apply our method on YouTube videos to obtain weakly-supervised data to improve per-frame methods. Note that we are only showing 5 out of 190 frames in the clip. Best viewed in colour on screen.

we are fitting SMPL body models [26] to the data, other information such as 2D keypoints and body-part segmentations can also be obtained automatically as done by [23].

By retraining the single-frame 3D pose estimator using our automatically-generated dataset, we obtain a more robust network that performs better on real-world (3DPW [52]) and mocap (HumanEVA [42]) datasets. We are thus the first paper, to our knowledge, to show how we can use masses of unlabelled real-world data to improve 3D pose estimation models.

2. Related Work

3D human pose is typically represented in the literature as either a point cloud of 3D joint positions or the parameters of a body model. A common approach with the former representation is to "lift" 2D keypoints (either ground truth or from a 2D pose detector) to 3D. This has been recently done with neural networks [28, 57, 31] and previously using a dictionary of 3D skeletons [38, 2, 59, 54] or other priors [47, 50, 2] to constrain the problem. The point cloud representation also allows one to train a CNN to regress directly from an image (instead of 2D keypoints) to 3D joints using supervision from motion capture datasets like Human 3.6M [35, 41, 34]. However, this approach overfits to the constrained environments of lab-captured motion capture datasets and does not generalise well to real-world images. Whilst methods based on "lifting" are more robust to this domain shift, they discard valuable information from the image as they depend solely on the input 2D keypoints.

Training models with supervision from both 2D keypoints (from real-world datasets such as [25, 3, 18]) and 3D joints (from mocap datasets) has been shown to help with generalisation to real-world images [58, 40, 29, 9, 44, 45]. However, greater success has been achieved in this scenario by fitting parametric models of human body meshes to images. Human body models, such as [26] and [5], encapsulate more prior knowledge, thus reducing the ambiguity of the 3D pose estimation problem. Explicit priors such as bone length ratios remaining constant [58, 9] and limbs being symmetric [9] are enforced naturally by body models. Moreover, this mesh representation also enables a direct mapping to body part segmentations [23, 36, 20].

Early work used the SCAPE body model [5] and fitted it to images using manually annotated keypoints and silhouettes [12, 42, 6, 14]. More recent works use the SMPL model [26] and fit it automatically. This is done by either solving an optimisation problem to fit the model to the data [7, 23, 55, 6] or by regressing the model parameters directly using a neural network [20, 32, 36, 49] or random forest [23]. Optimisation-based approaches minimise an energy function that depends on the reprojection error of the 3D joints onto 2D [7, 23], priors on joint angle and shape parameters [7, 23], and/or the discrepancy between the silhouette of the 3D model and its foreground mask in the 2D image [23, 6]. Direct regression methods, in contrast, train a neural network where the keypoint [20, 32, 36] or silhouette reprojection errors are used in its training objective [36, 32]. Kanazawa *et al*. [20] also use an adversarial loss that distinguishes between real and fake joint angles of SMPL models. This effectively acts as a joint-angle prior, allowing the au-

thors to utilise existing ground truth SMPL model fits from [27] without requiring them to be paired to images.

Our approach uses the per-frame neural network model of Kanazawa *et al*. [20] as the initialisation of our optimisation problem. Despite efforts by [20] to train it with realistic 2D data, we show (as illustrated by Fig. 1, 2) how this model often fails on challenging real-world videos and how these errors can be corrected with bundle adjustment. Moreover, we show how we can improve the performance of this network by finetuning it using the results of our bundle adjustment as ground truth on originally difficult sequences.

We note that despite there being previous efforts to produce temporally consistent fits of the SMPL model [16, 56, 55, 37], none of these works have been able to use these results to improve a per-frame model as we have. Furthermore, [56] and [37] have not explicitly evaluated on 3D pose estimation either. Additionally, we do not assume knowledge of calibrated cameras like [16, 55].

There are also several methods which enforce temporal consistency without body models: The works of [10, 53, 24] were based on Non-Rigid Structure from Motion whilst [4] lifted tracked 2D keypoints into 3D. More recently, Hossain *et al*. [15] also lifted 2D keypoints using an LSTM in a sequence-to-sequence [46] model. However, it is difficult to retain memory over long sequences as evidenced by their model performing best with a temporal context of only five frames. Dabral *et al*. [9] use a feedforward network using the predictions of the previous 20 frames as input. Our optimisation based approach, in contrast, can consider all frames (our experiments have as many as 1175 frames) in the video to produce more globally coherent results. Furthermore, as we consider all frames jointly, rather than sequentially like [9, 15], errors do not accumulate over time.

Finally, we note that there are several works which synthesise additional training data using rendering engines [39, 51, 8]. Although this approach provides additional diversity compared to motion capture datasets, the resultant data, although fully labelled, is not photorealistic. Our approach is complementary in that we leverage unlabelled, but real-world YouTube videos from the Kinetics dataset. Concurrently to this paper, [21] have also used additional videos from Instagram to improve 3D pose estimation models.

3. Bundle Adjustment using the SMPL Model

We jointly optimise the parameters of a SMPL statistical body shape model [26] and a camera over an entire video sequence. The whole-video approach contrasts with recurrent networks such as [15] which are only effective using a temporal context of only five frames, and allows for better global solutions. As shown in Fig. 2, the input to our method is a sequence of video frames, 2D keypoint predictions for a single person for each frame using a state-of-art 2D pose detector [33] and initial SMPL parameters produced per-frame using the HMR network of [20]. From this, our method outputs SMPL- and camera parameters for each frame in the video that are consistent with each other and reproject to the 2D keypoints. In Sec. 3.1, we briefly describe the SMPL body model that we are fitting to videos. Thereafter, in Sec. 3.2, we detail the objective function that we minimise in order to fit this model to the video. Section. 3.3 we provide details on the optimisation.

3.1. Body representation

The SMPL body model [26] parameterises a triangulated mesh with $N = 6890$ vertices of a human body. It factorises the 3D mesh into shape parameters, $\beta \in \mathbb{R}^{10}$ and pose $\theta \in \mathbb{R}^{3K}$, where $K = 23$ joints. The shape parameters model the variations in body proportions, height and weight. They are the coefficients of a low-dimensional shape space that was originally learned by [26, 7] from a training set of approximately 4000 registered human-body scans. The pose parameters model the deformation of the body as a result of the articulation of its K internal joints. They are an axis-angle representation of the relative rotation of a joint with respect to its parent in the model's kinematic tree. SMPL is a differentiable function that outputs a mesh and positions of joints in 3D. We denote the latter as $\mathbf{X} = \text{SMPL}(\beta, \theta) \in \mathbb{R}^{J\times 3}$ where J is the number of joints.

3.2. Formulation

We optimise an objective function that considers the reprojection of 3D keypoints onto 2D, temporal consistency of SMPL parameters, 3D- and 2D-keypoints, and a prior:

$$E(\beta, \theta, \mathbf{\Omega}) = E_R(\beta, \theta, \mathbf{\Omega}) + E_T(\beta, \theta, \mathbf{\Omega}) + E_P(\theta, \beta) \quad (1)$$

Reprojection error: We assume that we have 2D keypoint detections, $\mathbf{x}_{det,i}$ with a confidence score of w_i for the i^{th} joint. This error term penalises deviations of the projections of our estimated 3D joints onto 2D over all T frames in the video for all J body joints:

$$E_R(\beta, \theta, \mathbf{\Omega}) = \lambda_R \sum_t^T \sum_i^J w_i \rho(\mathbf{x}_i^t - \mathbf{x}_{det,i}^t). \quad (2)$$

Here, ρ is the robust Huber error function which we favour over a squared error as it can deal better with noisy estimates that we sometimes obtain on "in-the-wild" sequences, and the superscript t denotes time. $\mathbf{x}$ is the 2D projection of the 3D joint $\mathbf{X}$,

$$\mathbf{x}^t = s^t \Pi(R\mathbf{X}^t) + u^t \quad (3)$$

$$\mathbf{X}^t = \text{SMPL}(\beta, \theta^t), \quad (4)$$

where Π is an orthographic projection, $R \in \mathbb{R}^{3\times 3}$ is the global rotation matrix parameterised by a Rodrigues vector

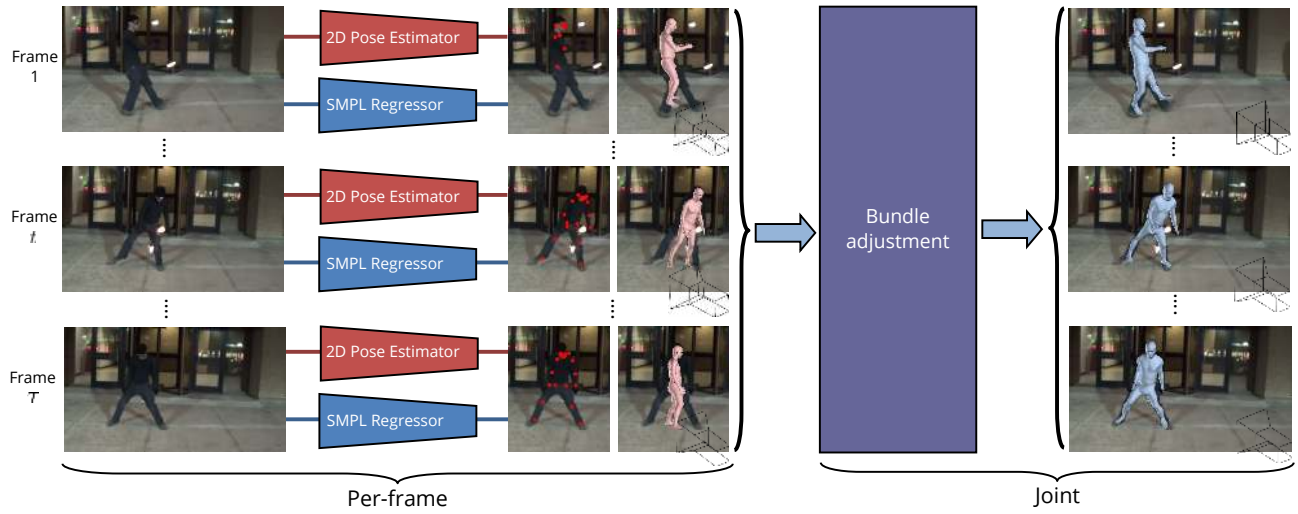

Figure 2. Overview of our method: Using initial per-frame estimates of 2D keypoints, SMPL- and camera parameters, we jointly optimise over the whole video comprising T frames by encouraging temporal consistency. As a result, we can overcome poor 2D keypoint detection (first row) and poor initial SMPL estimates (all rows) to output accurate SMPL- and camera-parameters.

and $\mathbf{\Omega}^t = \{s^t, u^t\}$ are the camera parameters comprising of scale, $s \in \mathbb{R}$ and translation $u \in \mathbb{R}^2$ and time-step t. Note that the parameters β and θ are mapped to 3D joint positions $\mathbf{X}$ by SMPL, and that we use a single β parameter for the whole sequence as the body shape of the video's subject remains constant.

Temporal error: This error, E_T is defined as:

$$E_T(\beta, \theta, \mathbf{\Omega}) = \sum_{t=2}^{T} \sum_{i=1}^{J} \lambda_1 \rho(\mathbf{X}_i^t - \mathbf{X}_i^{t-1}) + \lambda_2 \rho(\mathbf{x}_i^t - \mathbf{x}_i^{t-1}) + \lambda_3 \rho(\mathbf{\Omega}^t - \mathbf{\Omega}^{t-1}). \quad (5)$$

The temporal error on 3D joints, $\mathbf{X}$, and camera parameters, $\mathbf{\Omega}$, encourages smooth motions that are typical of humans in videos. This is also applied on the 2D keypoint projections, $\mathbf{x}$, as it helps to compensate for spurious errors of the 2D keypoint detector at a particular frame in the video.

3D Prior: There are many 3D poses (including some that are not humanly possible) that project correctly onto the 2D keypoints while also having low temporal error (for example, having all keypoints in a flat plane actually minimises the change with time). We use a single β for the entire sequence, meaning that changes in distance between 2D keypoints must be explained by pose changes, but telling which keypoint is in front of the other often remains ambiguous. Therefore, we include a prior term that encourages realistic 3D poses which match the appearance, as illustrated in Fig. 3. We use two terms: the same joint angle prior used by [7, 16, 23], and another term that robustly encourages the solution to stay close to our initialisation, $(\tilde{\beta}, \tilde{\theta})$, which was estimated by the single-frame HMR model. It is thus defined as:

$$E_P(\beta, \theta) = \sum_{t}^{T} E_J(\theta^t) + \lambda_I E_I(\theta^t, \beta) \quad (6)$$

$$E_J(\theta) = -\log \left(\sum_i g_i \mathcal{N}\left(\theta^t; \mu_i, \Sigma_i\right) \right) \quad (7)$$

$$E_I(\theta^t, \beta) = \sum_{i}^{J} \rho(\mathbf{X}_i^t - \tilde{\mathbf{X}}_i^t) + \lambda_\beta \rho(\beta - \tilde{\beta}^t). \quad (8)$$

The joint angle prior, $E_J(\theta)$, is the negative log-likelihood of a Gaussian Mixture Model that was fitted to the joint angles on the CMU Mocap dataset [1]. g_i are the mixture model weights of 8 Gaussians [7, 16, 23], and μ_i and Σ_i are the mean and covariance of the i^{th} Gaussian. Multiple modes are used to represent the diverse range of poses which a human can be in. Note that though our initialisation prior (8) penalises deviations in 3D joint positions, these are functions of the SMPL parameters according to (4).

3.3. Optimisation

We optimise (1) with respect to all SMPL and camera parameters, for all frames in the video, jointly using L-BFGS and Tensorflow. The solution is first initialised using the results of the per-frame, HMR neural network [20]. In total there are $10 + 75F$ parameters to be optimised for, where F is the number of frames in the video. On a typical clip from Kinetics [22] consisting of 250 frames, the optimisation takes about 8 minutes on a standard CPU or GPU (as we did not implement customised kernels for this task), or only 2 seconds per frame. The time- and memory-efficiency of our method is thus suited for batch, offline processing of videos as done in the following section.

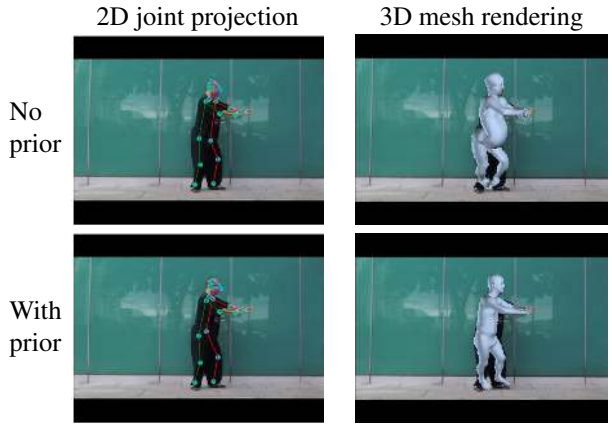

Figure 3. Without the prior (6), the SMPL model fit can project well onto 2D keypoints without being in a valid human pose.

3.4. Discussion

As previous works [37, 16, 55, 56, 30] have incorporated temporal information into 3D pose estimation using bundle adjustment before, we discuss the differences of our approach: First, in contrast to [37, 55, 30, 56], we use a robust Huber penalty function, and unlike previous approaches, also incorporate additional robustness into our reprojection term for Kinetics data in the next section. Second, our temporal consistency term is not only on 3D joint positions, but also on 2D joint projections and camera parameters (note that [16, 55, 30] assume known intrinsics). Third, unlike previous works, we use our bundle adjustment results to improve a per-frame model. Fourth, [37] optimises in the feature space of HMR, whilst we optimise SMPL- and camera-parameters directly. Additional segmentation masks for model fitting as also used by [16] and [55].

4. Leveraging Kinetics for weak supervision

Kinetics-400 [22] is a large-scale dataset of human actions collected from YouTube. It contains 400 or more 10s video clips for each of 400 action classes. Each clip is from a different YouTube video, and consequently the dataset contains considerable diversity in people, scenes and camera viewpoints as shown in Fig. 1,2,3. We perform bundle adjustment on this dataset to obtain real-world, weakly-supervised training data for 3D pose models. Bundle adjustment is challenging on Kinetics since there are often multiple people in a frame, shaking cameras, and people are often occluded or move off-camera. The diversity also results in more frequent failures of our multi-person 2D pose detector [33] and HMR [20].

Dealing with multiple people: We could handle multiple people with our formulation in Sec. 3 by first tracking a single person through the video, and applying our method to only the tracked region. However, we found this approach too sensitive to missing detections and tracking failures. Consequently, we perform tracking to initialise the solution but also augment the per-frame component of our loss function, (1), to deal with multiple (or potentially no) people, and allow for outliers to be ignored:

$$E_R(\beta, \theta^t, \mathbf{\Omega}^t; \mathbf{x}^t_{det,i}) = \min\left(\min_{p \in P^t} \sum_i^J w_i h(\mathbf{x}^t_i - \mathbf{x}^{t,p}_{det,i}), \tau_R\right), \tag{9}$$

$$E_I(\beta, \theta^t) = \min\left(\min_{p \in P^t} \sum_i^J \rho(\mathbf{X}^{t,p}_i - \tilde{\mathbf{X}}^t_i) + \lambda_\beta \rho(\beta - \tilde{\beta}^t), \tau_I\right). \tag{10}$$

Here, τ_R and τ_I are constants, and p indexes the different person detections P^t in frame t. Intuitively, the "inner min" means that the loss is with respect to the current best-matching 2D pose for each frame. However, if estimates from either the 2D pose detection or the HMR model are too far from the current bundle adjustment estimates, they are considered outliers, and the loss is set to a constant (performed by the "outer min"). This means that they no longer affect the bundle adjustment procedure. There is also substantial jitter in keypoint prediction in Kinetics, due to both 2D detector inaccuracy and camera shake. This causes significant problems if a bone is close to parallel with the camera plane: in such cases, jitter in 2D keypoints can often only be explained by large changes in 3D orientation. Since we are penalising 3D changes, this encourages the overall algorithm to avoid poses where bones are near parallel with the camera plane. To mitigate this, we replace the Huber loss, ρ, in the reprojection term with a hinge loss, h, which is 0 if the error is less than 5 pixels, and behaves like the Huber loss (i.e. L1 error) otherwise. Finally, to deal with camera motion, we find it advantageous to put an upper bound on the camera translations in (5), which is equal to 10% of the image width, and we do not penalise camera scaling.

Initialisation by tracking: The possibility of outliers means that initialisation is important, which we do by first tracking people in 2D using our multi-person pose detector [33] that outputs 2D keypoints and bounding boxes for each person in the image. We select bounding boxes by computing the shortest path from the start to the end of the video: distances between detected people in subsequent frames are equal to the mean-squared-error in pixels between detected keypoints. As there may be missing person detections, we allow the shortest-path algorithm to skip frames with a penalty of 100 pixels. Given a selected person detection for each frame, we initialise the 3D pose parameters for each frame using the estimates from HMR (for any skipped frames, we initialise using the pose from the nearest non-skipped frame).

Training data selection: After optimising, we measure the success of the algorithm by the total loss (1). However, we find that the loss tends to be lowest for people who aren't moving, producing videos that are not suitable to use as training data. This problem is alleviated by normalising the total loss by the the 3D trajectory length,

$$E_{norm}(\beta, \theta, \mathbf{\Omega}) = \frac{E(\beta, \theta, \mathbf{\Omega})}{\sum_t^T \sum_i^J \|\mathbf{X}_i^t - \mathbf{X}_i^{t-1}\|}. \quad (11)$$

To obtain training data, we process all videos in Kinetics that do not have more than 6 detected people in a single frame, as our 2D pose detector and HMR usually fail on crowded scenes. After running bundle adjustment, we then discard any videos where E_{norm} is above a threshold, retaining roughly 10% of the original videos. From these videos, we keep the frames where the 2D reprojections of the 3D poses are inliers with respect to our detected keypoints (i.e. $\min_{p \in P^t} \sum_i^J w_i \rho(\mathbf{x}_i^t - \mathbf{x}_{det,i}^{t,p}) < \tau_R$).

5. Experiments

After describing common experimental details in Sec. 5.1, we first analyse our bundle adjustment method on the Human 3.6M dataset in Sec. 5.2. Although this lab-captured dataset is not particularly realistic, it has metric ground truth 3D which allows us to conduct an ablation study and compare to previous work on 3D pose estimation using the SMPL model. Thereafter, in Sec. 5.3 we run our method large-scale on Kinetics videos before using these predictions in Sec. 5.4 as weakly-supervised ground truth to retrain a per-frame 3D pose estimation model as described previously in Sec. 4.

5.1. Experimental Set-up

We initialise the solution to bundle adjustment using the state-of-art HMR neural network [20] which is input an image and outputs SMPL and orthographic camera parameters. Unless otherwise specified, we use the publicly released model that has been trained on 3D mocap datasets: Human 3.6M [17] and MPI-3DHP [29], 2D pose datasets: COCO [25], MPII [3] and LSP [18], and an adversarial prior that was trained on SMPL model fits using [27]. The keypoints that we use for bundle adjustment are obtained using [33], which was trained on the same 2D pose data as HMR and additional data from Flickr collected by the authors.

5.2. Results on Human 3.6M

Human 3.6M [17] is a popular motion capture dataset and 3D pose benchmark. Following previous work [35, 40, 20], we downsample the videos from 50fps to 10fps and evaluate on the validation set. Even so, some videos contain as many as 1175 frames, which we are still able to jointly optimise over. We report the mean per joint position error (MPJPE) [17], and also this error after rigid alignment of the prediction with respect to the ground truth using Procrustes Analysis [11] which we denote as PA-MPJPE.

Table 1. Ablation study on Human 3.6M, considering the effect of different terms of our objective function (1). Mean errors over the validation set are reported.

Method	MPJPE (mm)	PA-MPJPE (mm)
HMR initialisation [20]	85.8	57.5
E_R	154.3	99.7
$E_R + E_P$	79.6	55.3
$E_R + E_P + E_T$	77.8	54.3
E_R (gt. keypoints)	89.2	64.5
$E_R + E_P$ (gt. keypoints)	66.5	45.7
$E_R + E_P + E_T$ (gt. keypoints)	63.3	41.6

Table 1 shows the effect of the various terms of our objective function in (1). We initialise the solution to our bundle adjustment using the public HMR model of [20], and the error increases if we only use the reprojection error. As shown in Fig. 3, optimising for reprojection error alone can result in impossible poses. Note that we are using a single β shape parameter across the whole video, but this alone is too weak a constraint. The addition of the prior term (6) improves results substantially: MPJPE reduces by 6.2mm compared to the HMR initialisation. Although HMR was also trained with 2D reprojection as one of its loss functions, we obtain better results by explicitly optimising for this term and using HMR as an initialisation method. Note that the 2D pose detector that we use [33] has not been trained on Human 3.6M at all. Our final model, which enforces temporal consistency with not only a single β parameter, but smoothness of joints and camera parameters, achieves the best results, significantly improving the MPJPE error of the initial HMR model by 9.4% and PA-MPJPE by 5.6%.

The final three rows of Tab. 1 use ground truth 2D keypoints. Note that here, as the ground truth is the projection of 3D joints into the image using the known camera, we have keypoints for occluded joints too. Each term of our objective function (1) has the same effect on the overall error as before. However, the MPJPE and PA-MPJPE improve considerably more over the initialisation of HMR: Our final model reduces these errors by 26.2 and 27.2% respectively. This shows the significant benefits that we can obtain if we have knowledge of occluded keypoints since this further reduces the ambiguity in the problem.

Finally, Tab. 2 shows we achieve the best results on Human 3.6M among other methods utilising the SMPL model. Note that Mehta *et al.* [30] also perform bundle adjustment to improve the predictions of a CNN model, obtaining an MPJPE of 80.5. However, as [30] do not use the SMPL model, they are not directly comparable. Additionally, although direct CNN-regression methods such as [9] obtain MPJPE errors of as low as 52.1, they overfit to the Human 3.6M dataset and have been shown to be signifi-

Table 2. Comparison of approaches fitting the SMPL model [26] on Human 3.6M. We did not use additional Kinetics data here.

Method	MPJPE (mm)	PA-MPJPE (mm)
Self-Sup [49]	–	98.4
Lassner *et al*. direct fitting [23]	–	93.9
SMPLify [7]	–	82.3
Lassner *et al*. optimisation [23]	–	80.7
Pavlakos *et al*. [36]	–	75.9
NBF [32]	–	59.9
MuVS (Note uses 4 cameras) [16]	–	58.4
HMR [20]	88.0	56.8
Ours	**77.8**	**54.3**

Table 3. Statistics of our bundle-adjustment dataset from Kinetics-400. 2D inliers refers to frames where 2D reprojection error was small: $E_R < \tau_R$.

	Count
Total videos	106 589
Selected videos ($E_{norm} < \tau_R$)	15 046
Total frames in selected videos	3 730 672
BA inliers	3 045 603

cantly outperformed by SMPL-based approaches on real-world datasets such as 3DPW [52] by Kanazawa *et al*. [21].

5.3. Results on Kinetics

Given that our algorithm can reliably improve 3D estimates, we apply our method to a large-scale video dataset to produce training data for single-frame 3D pose estimation. We used the entirety of Kinetics-400 [22] (400+ clips of 400 action classes), after automatically selecting videos as described in Sec. 4.

Table 3 shows the statistics of the important stages in this process. We first pre-select roughly 15K videos based on the normalized bundle adjustment loss (11), resulting in 3.7M frames. The bundle adjustment matched the prediction of the 2D pose detector [33] for 3.0M out of 3.7M frames (we used a threshold of $\tau_R = 50$ pixels total error to determine outliers). Visual inspection showed that the 3D pose detector was fairly reliable: for the majority of outlier frames, the person was occluded or had simply left the frame.

Table 4 lists the action classes from Kinetics that were selected most often, showing that none of them appear in existing mocap datasets [17, 43, 29]. Mocap datasets only contain actions performed by a single person, in contrast to classes such as "tap dancing" and "salsa dancing" which bundle adjustment performs well on. Similarly, our method is effective on outdoor activities such as "roller skating" and "playing tennis" which cannot be recorded by mocap. There were no classes without any selected videos, but for several classes (e.g., "scrambling eggs" and "tying tie"), where a person is rarely fully visible, we only selected 1 video each. Some qualitative examples of the diversity of our dataset are shown in Fig. 4. All experimental hyperparameters are included in the arxiv version.

Table 4. The most common action classes of the videos selected from Kinetics. Our bundle adjustment method works well on action classes that do not appear in motion capture datasets, e.g., those that occur outdoors or contain multiple people.

Action class	Selected videos	Selected frames
Hula hooping	237	55 481
Roller skating	237	52 616
Spinning poi	210	40 251
Dribbling basketball	201	42 136
Playing tennis	191	45 824
Salsa dancing	187	42 744
Tap dancing	187	43 947

Table 5. Results on the 3DPW [52] and HumanEVA [43] datasets when training with our Kinetics datasets. We evaluate the HMR model retrained by us on its original training data using the author's public code, and the HMR model trained on its original data and 300K and 3M frames from our Kinetics dataset. We report the PA-MPJPE error in mm.

Dataset	Original data	Original + Kinetics 300K	Original + Kinetics 3M
3DPW	77.2	73.8	**72.2**
HumanEVA	85.7	83.5	**82.1**

5.4. Weak supervision from Kinetics

We utilise the training data that we automatically obtained in the previous section to retrain a new HMR model from Imagenet initialisation. We use the original training data (described in Sec. 5.1) too, and use a model only trained on this data as our baseline. We evaluate on the recently released 3D Poses in the Wild dataset (3DPW) [52], which consists of outdoor videos captured in real-world conditions and HumanEVA [43], a mocap dataset. Our network has never been trained on images from either dataset. To verify the effect of Kinetics training, we trained a model with all frames from our automatically-generated dataset (Kinetics 3M), and also with a random subset of 10% of the frames in our dataset (Kinetics 300K).

When retraining the HMR model on Kinetics data, we made modifications to the HMR training procedure [20]. These are detailed in the arxiv version, where we also show that our modifications only help for training on Kinetics data, and not when using only the original training data used by HMR.

3D Poses in the Wild: This recently released dataset contains 60 clips, consisting of outdoor videos captured from a moving mobile phone and 17 IMUs attached to the subjects [52]. The IMU data allowed the authors to accurately compute 3D poses which we use as ground truth. We evaluate on the test set comprising 24 videos; details of our evaluation protocol are in the arxiv version.

Table 5 shows how using additional data from Kinetics improves results on this dataset. Training with 300K frames of Kinetics data improves the PA-MPJPE by 3.4mm,

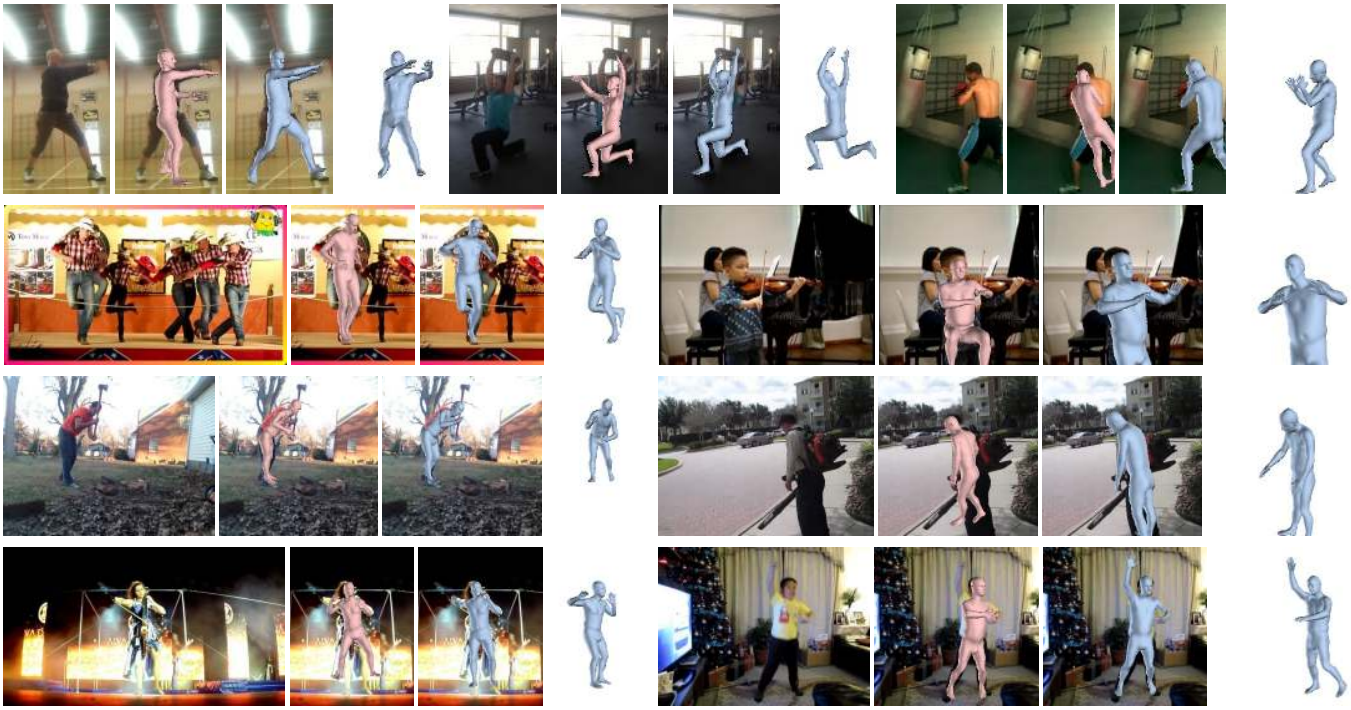

Figure 4. The dataset we automatically generated from Kinetics has a diverse range of scenes, people, camera viewpoints and action classes not found in motion capture. We show the input frame, results for a single tracked person (which are cropped for display) of HMR (pink) and bundle adjustment (blue), and the bundle adjustment result from another view respectively. Note how bundle adjustment typically improves the per-frame estimates of HMR.

and our model trained with all 3M frames of Kinetics improves further by 5 mm over the baseline. Our Kinetics-trained model also outperforms the public HMR model [20] (trained by the authors) which obtains a PA-MPJPE error of 74.9. While isolated checkpoints from our reimplementation of HMR perform as well as the public model, not all do; Tab 5 computes the mean of 20 checkpoints (roughly 1500 training iterations apart) to minimise variance.

HumanEVA: HumanEVA [43] is an indoor motion-capture dataset where we follow the evaluation protocol of [7] on the validation set. Although HumanEVA does not contain "in the wild" data, it is a dataset which our HMR model has not been trained on at all. Table 5 shows how adding additional data from our Kinetics dataset improves performance on this dataset compared to our baselines that were trained without Kinetics. Our model trained with 300K frames of Kinetics data improves the PA-MPJPE by 2.2 mm, and the model trained with 3M Kinetics frames improves further by 3.6 mm over our baseline. The public HMR model obtains a PA-MPJPE error of 83.5, which is also worse than our Kinetics-trained model.

These experiments thus show how we can effectively use Kinetics data to improve the per-frame HMR model on multiple datasets. We also achieve greater improvements on the real-world 3DPW dataset, compared to the mocap HumanEVA dataset.

6. Conclusion and Future Work

We presented a bundle-adjustment algorithm to leverage the temporal context in a video in order to improve estimates of the 3D pose of a person. Furthermore, we applied this to YouTube videos from Kinetics and automatically generated a dataset which we used to improve per-frame 3D pose estimators, demonstrating how we can effectively use large amounts of unlabelled data to improve existing models.

Bundle adjustment was effective because videos are shot in a 3D world where people move slowly (relative to the camera framerate), and the person's size and appearance remain consistent over time. If properly characterised, these constraints can give strong supervision to algorithms, which allows us to break out of the environments which motion capture devices are restricted to. We believe there is far more 3D structure to exploit, because people don't behave in a vacuum. People act under gravity, are supported by ground planes and interact with objects. Therefore, we aim in future to use physical constraints and information about human actions to constrain poses and predict the objects that people are interacting with to estimate their affordances.

Acknowledgements: We thank Jean-Baptiste Alayrac, Relja Arandjelović, João Carreira, Rohit Girdhar, Viorica Pătrăucean and Jacob Walker for valuable discussions.

References

[1] CMU graphics lab motion capture database. http://mocap.cs.cmu.edu/. 4

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

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

[4] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, pages 623–630, 2010. 3

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

[6] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *CVPR*, pages 1–8, 2007. 2

[7] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 2, 3, 4, 7, 8

[8] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *3DV*, 2016. 3

[9] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, and A. Jain. Structure-aware and temporally coherent 3D human pose estimation. In *ECCV*, 2018. 2, 3, 6

[10] P. F. Gotardo and A. M. Martinez. Non-Rigid Structure from Motion with complementary rank-3 spaces. In *CVPR*, 2011. 3

[11] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 6

[12] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *ICCV*, pages 1381–1388. IEEE, 2009. 2

[13] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 1

[14] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H.-P. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *CVPR*, pages 1823–1830, 2010. 2

[15] M. R. I. Hossain and J. J. Little. Exploiting temporal information for 3D pose estimation. In *ECCV*, 2018. 1, 3

[16] Y. Huang, F. Bogo, C. Classner, A. Kanazawa, P. V. Gehler, I. Akhter, and M. J. Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, pages 421–430, 2017. 3, 4, 5, 7

[17] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014. 1, 6, 7

[18] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. doi:10.5244/C.24.12. 2, 6

[19] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 1

[20] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[21] A. Kanazawa, J. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 3, 7

[22] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. In *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 4, 5, 7

[23] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 2, 4, 7

[24] X. Li, H. Li, H. Joo, Y. Liu, and Y. Sheikh. Structure from recurrent motion: From rigidity to recurrency. In *CVPR*, pages 3032–3040, 2018. 3

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2, 6

[26] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6):248:1–248:16, Oct. 2015. 1, 2, 3, 7

[27] M. M. Loper, N. Mahmood, and M. J. Black. MoSh: Motion and shape capture from sparse markers. *ACM TOG*, 33(6):220:1–220:13, Nov. 2014. 3, 6

[28] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 1, 2

[29] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017. 1, 2, 6, 7

[30] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM TOG*, 36(4):44, 2017. 5, 6

[31] F. Moreno-Noguer. 3D human pose estimation from a single image via distance matrix regression. In *CVPR*, pages 1561–1570. IEEE, 2017. 2

[32] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *3DV*, 2018. 2, 7

[33] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017. 3, 5, 6, 7

[34] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *CVPR*, 2018. 2

[35] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, pages 1263–1272, 2017. 2, 6

[36] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 1, 2, 7

[37] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine. Sfv: Reinforcement learning of physical skills from videos. *arXiv preprint arXiv:1810.03599*, 2018. 3, 5

[38] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3D human pose from 2D image landmarks. In *ECCV*, pages 573–586. Springer, 2012. 2

[39] G. Rogez and C. Schmid. Image-based synthesis for deep 3D human pose estimation. *IJCV*, pages 1–16, 2018. 3

[40] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 2, 6

[41] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe. How robust is 3D human pose estimation to occlusion? In *arXiv preprint arXiv:1808.09316*, 2018. 2

[42] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, pages 1337–1344, 2008. 2

[43] L. Sigal, A. O. Balan, and M. J. Black. HumanEVA: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4, 2010. 1, 7, 8

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

[45] X. Sun, B. Xiao, S. Liang, and Y. Wei. Integral human pose regression. In *ECCV*, 2018. 1, 2

[46] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014. 3

[47] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *CVIU*, 80(3):349–363, 2000. 2

[48] W. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment: A modern synthesis. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, 2000. 1

[49] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, pages 5236–5246, 2017. 2, 7

[50] J. Valmadre and S. Lucey. Deterministic 3D human pose estimation using rigid structure. In *ECCV*, 2010. 2

[51] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017. 3

[52] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2, 7

[53] B. Wandt, H. Ackermann, and B. Rosenhahn. 3D reconstruction of human motion from monocular image sequences. *PAMI*, 38(8):1505–1516, 2016. 3

[54] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3D human poses from a single image. In *CVPR*, pages 2361–2368, 2014. 2

[55] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes–the importance of multiple scene constraints. In *CVPR*, 2018. 2, 3, 5

[56] X. Zhang, T. Dekel, T. Xue, A. Owens, Q. He, J. Wu, S. Mueller, and W. T. Freeman. Mosculp: Interactive visualization of shape and time. In *arXiv preprint arXiv:1809.05491*, 2018. 3, 5

[57] R. Zhao, Y. Wang, and A. M. Martinez. A simple, fast and highly-accurate algorithm to recover 3D shape from 2D landmarks on a single image. *PAMI*, 2017. 2

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

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