H3WB: Human3.6M 3D WholeBody Dataset and Benchmark

Yue Zhu, Nermin Samet, David Picard

H3WB: Human3.6M 3D WholeBody Dataset and Benchmark

Yue Zhu Nermin Samet David Picard
LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France
{yue.zhu, nermin.samet, david.picard}@enpc.fr

Code and dataset: https://github.com/wholebody3d/wholebody3d

Abstract

We present a benchmark for 3D human whole-body pose estimation, which involves identifying accurate 3D keypoints on the entire human body, including face, hands, body, and feet. Currently, the lack of a fully annotated and accurate 3D whole-body dataset results in deep networks being trained separately on specific body parts, which are combined during inference. Or they rely on pseudo-groundtruth provided by parametric body models which are not as accurate as detection based methods. To overcome these issues, we introduce the Human3.6M 3D WholeBody (H3WB) dataset, which provides whole-body annotations for the Human3.6M dataset using the COCO Wholebody layout. H3WB comprises 133 whole-body keypoint annotations on 100K images, made possible by our new multi-view pipeline. We also propose three tasks: i) 3D whole-body pose lifting from 2D complete whole-body pose, ii) 3D whole-body pose lifting from 2D incomplete whole-body pose, and iii) 3D whole-body pose estimation from a single RGB image. Additionally, we report several baselines from popular methods for these tasks. Furthermore, we also provide automated 3D whole-body annotations of TotalCapture and experimentally show that when used with H3WB it helps to improve the performance.

1. Introduction

3D Human pose estimation is the task of localizing human body keypoints in images which is critical to analyze human behavior, expressions, emotions, intentions, and how people communicate and interact with the physical world. As a result, 3D human pose estimation has an important role in several vision tasks and applications such as robotics [28, 25, 29] or augmented/virtual reality [55, 3, 73, 81]. However, to make more accurate predictions about human behaviors, we need more than a few body keypoints. To that end, 3D whole-body pose estimation aims to detect face, hand and foot keypoints in addition to the standard human body keypoints of classical 3D human pose estimation.

The lack of accurate 3D datasets has made 3D whole-body pose estimation a challenging task, leading previous works to focus on separate body parts and train separate models on different datasets for 3D body pose [1, 30, 55, 56, 63, 3, 44, 45, 46, 65, 74, 77, 80, 26], 3D hand pose [8, 58, 84, 7, 27, 37, 82, 31], or 3D face landmarks [68, 9, 13]. However, directly ensembling separate body part models during inference suffers from issues arising from datasets' biases, pose and scales, and complex inference pipelines. Distillation from pretrained models has been used to overcome these issues, with FrankMocap [66] using three specialized pretrained models to estimate 65 whole-body keypoints (22 on the body, 40 on the hands, and 3 on the face), and DOPE [75] also using three specialized models to output 139 whole-body keypoints (13 on the body, 42 on the hands, and 84 on the face).

Alternatively, parametric body models can be fitted to obtain whole-body pose, as has been proposed in ExPose [18], SMPLify-X [60] or Monocular Total Capture (MTC) [77]. While parametric models enable sampling an almost infinite number of keypoints from the mesh [24, 22, 23], their accuracy is usually less than that of detection based methods on fine body parts like hands and feet (see supplementary material for examples from the literature). Indeed, parametric models are tailored for visual applications such as realistic motion generation [62] or avatar capture [67], where a realistic capture of the body shape and pose is more important than fine grain accuracy. Relying on a small set of data-driven pose parameters ensures realism but also limits their flexibility in representing complex unusual poses. For applications where the body shape is not needed but the accuracy of the keypoints is essential, like high performance sport analysis or ergonomics, detection based methods are thus the preferred solution.

Furthermore, 3D whole-body pose estimation has not been fully explored in the literature due to the absence of a representative and accurate benchmark. As previously mentioned, existing 3D whole-body methods either rely on specific datasets and models for different body parts, lead-

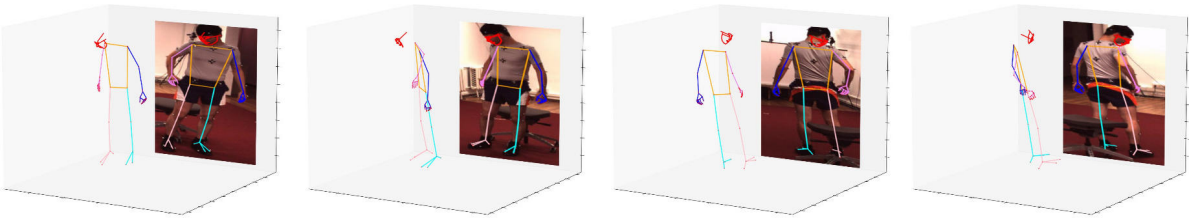

Figure 1. The H3WB dataset has 133 whole-body keypoint annotations in 3D as well as their respective projections in 2D.

ing to complex training pipelines and heterogeneous evaluations, or utilize parametric models that prioritize shape capture over highly precise keypoints. In addition, unified methods vary significantly in terms of keypoint layout definition, number of keypoints and distribution of keypoints across body parts (see Table 1). These significant dataset disparities and the absence of a standard benchmark make it challenging to compare methods fairly.

To address the above issues, we propose a new large-scale dataset for accurate 3D whole-body pose estimation called Human3.6M 3D WholeBody, or H3WB for short (see Figure 1). Our dataset extends Human3.6M [36, 12] with 3D whole-body keypoint annotations. It consists of 133 paired 2D and 3D whole-body keypoint annotations for a set of 100k images from Human3.6M, following the same layout used in COCO WholeBody [40]. More specifically, in addition to the standard 17 body keypoints, the dataset has 42 hand keypoints, 6 foot keypoints and 68 facial landmarks. H3WB was automatically created in a 3 step process: We obtain an initial set of 3D annotations using multi-view geometry. Then, we trained a masked auto-encoder to complete the initial annotations. Finally, we refine the whole-body keypoints via a diffusion model. A manual annotation of 80K keypoints from 600 images shows our labels have an average error of 17mm which suggest the H3WB keypoints are very accurate for such a complex task. We propose 3 tasks and benchmarks on the H3WB dataset for which we provide baselines: i) *3D whole-body pose lifting from a complete 2D whole-body keypoints*, ii) *3D whole-body pose lifting from incomplete 2D whole-body keypoints* (i.e. 2D whole-body with missing keypoints, which is more realistic), and iii) *3D whole-body pose estimation from a single RGB image*.

Our contributions can be summarized as follows. 1) We propose a method to create detailed 3D human pose keypoints from multi-view images. 2) We propose H3WB, the first accurate public benchmark dataset for 3D whole-body pose estimation, using the aforementioned method. Our benchmark can easily leverage existing results in 2D and enables the community to build upon existing high-quality 2D detectors on COCO. Unifying the 3D whole-body pose estimation with the COCO 2D benchmark will greatly benefit the research community. 3) We provide baselines for the 3 tasks of H3WB, which we believe will encourage the community to explore 3D whole-body pose estimation more and accelerate progress in the field. 4) Additionally, we provide 3D whole-body annotations for the TotalCapture [43] dataset, and show that when combined with the H3WB dataset it improves the performance of pose lifting tasks.

2. Related work

3D Body, hand and face pose estimation. There are two main groups of prominent approaches in 3D human pose estimation. The first group directly estimates 3D body pose from a single RGB image [61, 54, 55, 56, 61, 63]. The second group follows two stage approach where they first localize 2D keypoints and then lift 2D human pose to 3D space [53, 72, 38, 6, 49, 53]. Several optimization based methods [6, 49], utilize 2D keypoints to initialize a parametric model of the human body such as SMPL [51]. Several works attempt to eliminate the requirement of 3D annotations using 2D multi-view supervision to estimate 3D human pose [72, 78] or temporal supervision with video [17, 50]. 3D hand pose estimation methods share similar approaches as the body counterparts. First group of works, estimates hand pose from a single RGB image by directly regressing 3D hand keypoints [79], mesh vertices [27, 48], and parameters of parametric 3D hand models such as MANO [7, 64, 2, 14, 15, 82]. Second groups of works rely on intermediate 2D representations such as 2D keypoints and feature maps [8, 58, 82, 84, 34]. Simi-

Dataset	Size	Keypoints	Body	Hand	Face
Human3.6M[36]	3.6M	17	17		
3DPW[71]	51k	24	24		
LSP[41]	10k	14	14		
3DHP[71]	>1.3M	17	17		
Panoptic[42]	1.5M	15	15		
MTC[77]	834K	20	20		
InterHand2.6M[57]	2.6M	21		21	
FreiHAND[85]	37k	21		21	
RHD[84]	44K	21		21	
MTC[77]	111K	21		21	
TotalCapture[43]	1.9M	127	21	16+16	74
ExPose[18]	33K	144	25	15+15	89
H3WB	100k	133	23	21+21	68

Table 1. Overview of datasets for 3D human pose estimation.

larly, predominant 3D face pose estimation methods regress the dense 3D face landmarks [19, 21, 39] and face model parameters [13, 68, 69, 20, 76] based on 3DMM [5].

3D Whole-body pose estimation. There are a several methods [60, 77, 83, 18, 75, 66] jointly estimating 3D whole-body pose. The first group of works is based on parametric human body models such as Adam [43] and SMPL-X [60]. MTC [77] is based on the Adam model [43], and first gets 2.5D predictions, then optimizes the parameters of Adam. SMPLify-X optimizes the parameter of the SMPL-X model [60] to fit it to 2D keypoints. As a major drawback, optimization-based methods are relatively slow and highly sensitive to parameter initializations. Non-parametric methods [75, 18, 66] follow different approaches to avoid heavy optimization procedure. DOPE [75] and FrankMocap [66] first train separate body, hand, and face models. Next, they combine those models within a learning framework. DOPE [75] curates pseudo-ground truths from separate body models and uses those ground-truths to supervise the distillation model. Similar to DOPE, Ex-Pose [18] first obtains a pseudo-ground truth dataset by fitting SMPL-X model on in-the-wild images, and trains a joint model to output whole-body poses. All these methods utilize many part-based datasets. Moreover, all output different whole-body layouts with a different numbers of whole-body keypoint. FrankMocap, DOPE, and SMPLify-X estimate whole-body pose with 65, 139 and 144 keypoints, respectively.

Pose completion completes a partially estimated pose by localizing missing keypoints. Carissimi *et al*. [10] propose a denoising variational autoencoder network to fill the missing keypoints in 2D pose completion. Bautembach *et al*. [4] selects a small subset of poses from a database based on their distance to an incomplete 3D pose, and replaces the missing keypoints with the corresponding averaged keypoints in the subset. Despite being critical for real-world scenarios, pose completion has not been sufficiently explored due to the lack of annotated keypoint datasets. Our 3D whole-body dataset can facilitate more exploration of 3D pose estimation from 2D incomplete human poses.

3. The H3WB dataset

In this section, we describe the making of the H3WB dataset[1]. Our objective is to build a keypoint based 3D whole-body dataset including keypoints on the body, the face and the hands, and propose a benchmark. We use the same keypoint layout as COCO WholeBody dataset [40] with 133 keypoints. To that end, we build on the widely used Human3.6M dataset [36] for which we provide 3D whole-body keypoints. The H3WB building process is as follows: First, we use an off-the-shelf 2D whole-body detector combined with multi-view reconstruction to obtain an initial set of incomplete 3D whole-body keypoints. Next, we implement a completion network to fill in the keypoints missed by the multi-view geometric approach. Then, we develop a refinement method for the hands and the face to obtain more accurate keypoints. Finally, we perform quality assessment to select 25k 3D whole-body poses with high confidence and the 100k associated images from 4-view.

[1]We consider the feet keypoints as a part of the body.

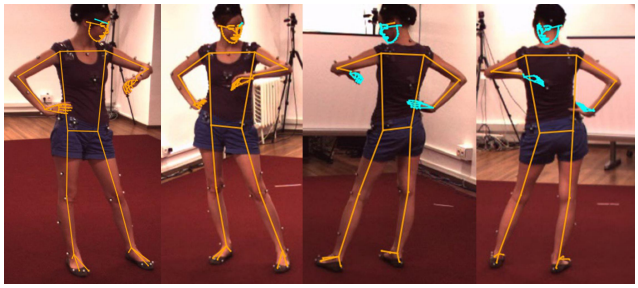

Figure 2. OpenPifPaf detects most of the non-occluded keypoints inside the image (orange keypoints). The occluded or undetected keypoints (cyan keypoints) are reprojections after 3D multi-view reconstruction. Notice that these reprojections do not always align with the images, like the right hand in the last view, which is probably due to OpenPifPaf not being perfectly accurate.

3.1. Initial 3D whole-body dataset with OpenPifPaf

We run the 2D whole-body detector from OpenPifPaf [47] on all the 4 views from the training set of Human3.6M (S1, S5, S6, S7 and S8, 1 image per 5 frames). Since the cameras of Human3.6M are well calibrated, we can reconstruct keypoints in 3D using standard multi-view geometry.

The OpenPifPaf 2D whole-body detector can miss keypoints due to self-occlusions (hands, feet) or unfavorable camera viewpoints (facial landmarks). However, the four-view setup allows us to recover missing keypoints and obtain a complete 3D whole-body pose, provided each keypoint appears in at least two non-opposing views. An example of this process is shown in Figure 2. Using this method, we obtained 11,426 fully complete 3D whole-body poses with all 133 keypoints and 26,333 incomplete 3D whole-body poses where all keypoints appear in at least one view, resulting in a total of 37,759 3D whole-body poses with each keypoint appearing in at least one view.

We did not rely of the video information because the reconstruction problem becomes significantly more difficult when there is a motion between two frames. In the absence of motion, an additional frame does not help solve the occlusion problem. The results of our study demonstrate that multi-view labeling is sufficiently effective for our task (see Table 2, "Geometry" line).

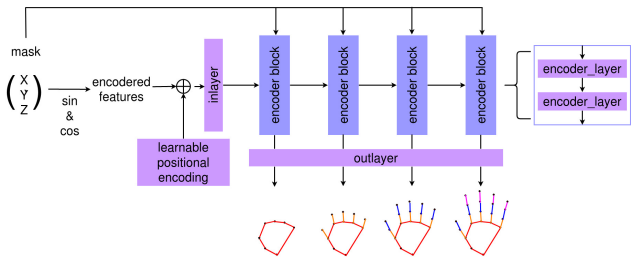

Figure 3. The completion network consists of one linear input layer, 4 transformer encoder blocks (each of them containing 2 transformer encoder layer with $d\_model = 64$ and $n\_head = 1$), and a linear output layer. At the end of each encoder block, the features are decoded by the output layer into a predicted position in a curriculum way where later blocks decode more keypoints.

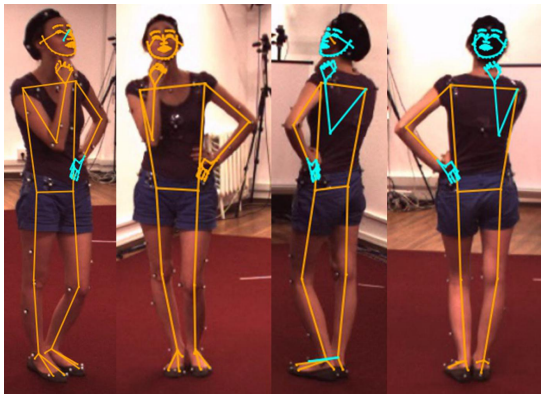

Figure 4. Example outputs of the completion network. The orange color denotes the keypoints that were detected by OpenPifPaf. The cyan color shows the missed keypoints by OpenPifPaf but completed by our completion network. The left hand is detected in only 1 view by OpenPifPaf and thus fully predicted by the completion network.

3.2. Completion network

In order to complete the 26,333 incomplete 3D whole-body poses, we develop a completion network as shown in Figure 3. We designed our completion network using Transformer architecture [70] as they can easily handle the conditional dependencies introduced by the skeleton's topology through masking. Since each skeleton always has exactly 133 keypoints, which can be considered as 133 tokens of 3 coordinate values. Token values are expanded from 3 coordinates to $3 \times 16 = 48$ features using Fourier encoding. We use learnable positional encoding since each keypoint is uniquely identified.

We train the completion network on the 11,426 complete skeletons using a masked auto-encoder strategy [32] where the missing keypoints are masked at the input and will be predicted using the unmasked keypoints. The masking strategy is as follows:

- With a 50% chance, we perform a keypoint wise mask where each keypoint has 15% chance of being masked,
- with the remaining 50% chance, we perform a block wise mask in which either the body, the left hand, the right hand, the left or the right part of the face are masked (uniform probability).

To ease the learning process and take into account the causal link between some keypoints (*e.g.*, the tip of a finger depends on the position of its parent phalanges), we introduce a curriculum approach. We compute the loss at different levels following a hierarchy where early levels consider only keypoints closer to the root, while later levels consider more deformable keypoints which highly depend on their parents. We illustrate the completion network and learning process in Figure 3. The loss function is

$$\begin{aligned}\mathcal{L}(X, X_{gt3D}, X_{gt2D}) =& \mathcal{L}_{3D}(X, X_{gt3D}) \\ &+\alpha\mathcal{L}_{2D}(X, X_{gt2D}) \\ &+\beta\mathcal{L}_{sym}(X), \end{aligned} \tag{1}$$

where $\mathcal{L}_3D$ is an ℓ_1 loss of 3D coordinates, $\mathcal{L}_2D$ is an ℓ_1 loss of 2D projection of the 3D coordinates if we have the 2D annotation from OpenPifPaf, and $\mathcal{L}_{sym}$ is a symmetric loss which is applied to make sure the left part and right part of the human have the same length on corresponding body parts.

We show an example output from our completion network in Figure 4. The completion network results on missing body parts are visually realistic and appealing. However, since the completion network does not rely on the image content, its output does not always align with the image

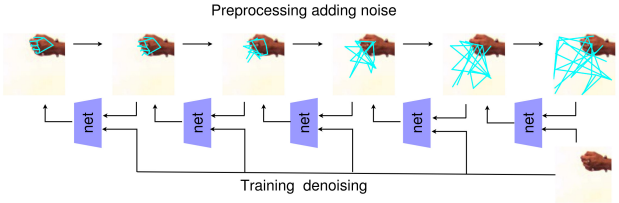

Figure 5. Refinement network architecture and training process. Gaussian noise is added to the groundtruth coordinates with increasing variance, and the network is iteratively trained to recover the less noisy coordinates.

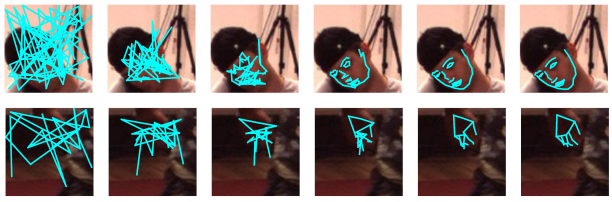

Figure 6. Example outputs from the face (top row) and hand (bottom row) refinement networks during inference time. We observe that the predictions almost converge to the correct locations in 5-iteration.

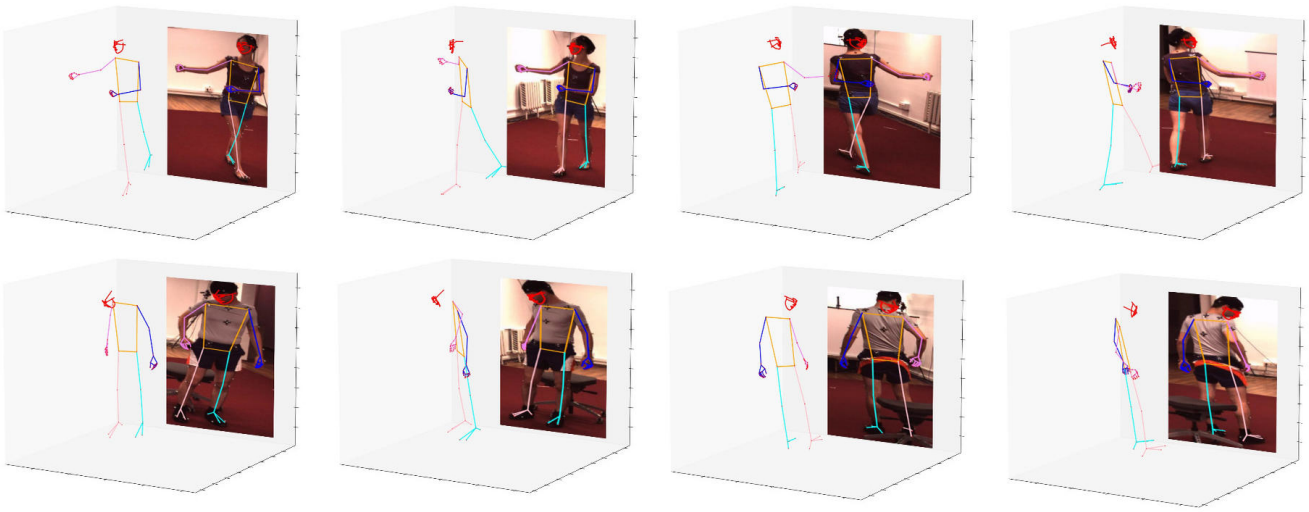

Figure 7. Examples of the 3D whole body skeleton. They are visually realistic humans. The strange looking faces (fatter or thinner) in different views are due to viewing artifacts of the default perspective projection.

and may only reflect the most common poses of the training set. This can quantitatively be seen in the line "+ Completion" of Table 2.

3.3. Hands and face 2D refinements

In order to correct the alignment problem, we propose another neural network that refines the 2D position of keypoints on the face and the hands. Previous studies have explored and demonstrated the effectiveness of 2D human pose refinement using an iterative error feedback framework [11]. Motivated by this, we build upon recent conditional diffusion models [35] and we consider the prediction from the completion network as *noisy* such that the refinement network *denoises* it to conditionally fit the image.

We train separate refinement models for the face and the hands, while keeping the same network architecture and the same training strategy. We used a simple MLP and found it to be effective, preventing the need to explore more complex architectures. We illustrate the refinement process in Figure 5. During training, we add Gaussian noise to the groundtruth poses with an increasing variance from 5 to 25 pixels, and annotate them as step $t = 1...5$ (step $t = 0$ is the groundtruth). The network learns to predict the pose at step t given the image and the noisier step $t + 1$ with a 2D supervision loss.

We build two small datasets, each consisting of 22,000 non-occluded faces and hands respectively, with their corresponding OpenPifPaf predictions. Each image is resized to 384×384 pixels. We use a random crop of size 224×224 pixels to have the face and hands located in diverse regions of the images. We split the datasets into training and validation sets with 20,000 images and 2,000 images, respectively.

Quantitatively, the face predictions achieve an average error less than 3 pixels and the hand predictions achieve an average error less than 7 pixels on the validation sets. We show example qualitative results in Figure 6.

Finally, we run the refinement networks on the 2D-projections of the 3D poses predicted by our completion network. For each 3D skeleton, we project it into the 4 different 2D views. We then crop the regions around the hands and face and denoise the corresponding predictions using the refinement network with 10 iterations to obtain refined 2D poses in each of the 4 views.

Although the refinement network is not always correct due to its training on non-occluded faces or hands, we only need 2 non-opposing views to perform geometric reconstruction. Since bad refinements tend to collapse all keypoints into the same location, we select the two non-opposing views with the highest variance in keypoint positions to avoid disruptions caused by occlusions. Using this method, we obtain 151,036 triplets of 3D whole-body keypoints, corresponding image, and 2D projected keypoints from the original set. Examples of resulting 3D whole-body skeletons and their image-aligned 2D counterparts are shown in Figure 7 and Figure 8, respectively.

3.4. Quality assessment

To select the most accurate triplets from our dataset, we reuse the refinement networks and employ a multi-crop strategy that accounts for the variance of the prediction. We project each 3D whole-body skeleton onto all 4 views, and produce four cropped images for each region of interest around the face and hands. The refinement network is run on these 4 crops, and the resulting predictions are aligned with the original prediction to compute the 2D error compared to the original 2D projection. We score the

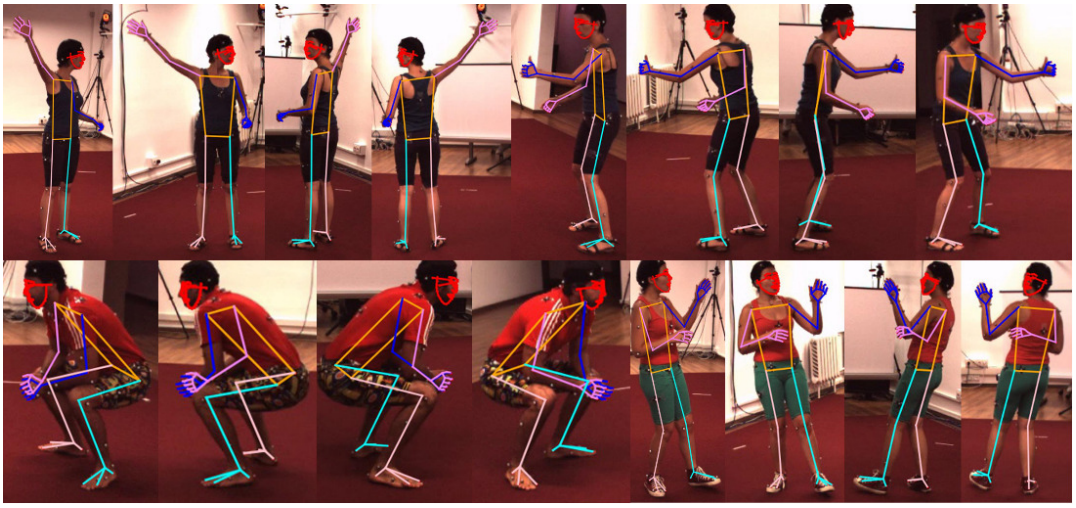

Figure 8. Examples of the 3D whole-body skeleton projected in 2D onto their corresponding images. They are visually accurate, though still there are small errors in detail which we do not expect to overcome due to the initial resolution and ambiguity of the images.

3D skeletons by averaging the errors of all 4 projected views, and select the 5k lowest error skeletons from each subject of Human3.6M (S1, S5, S6, S7, S8) to form the $5\text{k} \times 4(\text{view}) \times 5(\text{subject}) = 100\text{k}$ triplets of {image, 2D coordinates, 3D coordinates in camera space} of our 3D whole-body dataset.

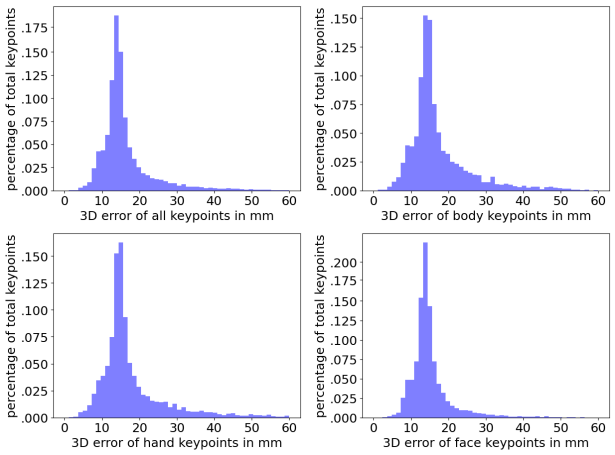

Figure 9. 3D error distributions calculated from 80k manually corrected keypoint annotations. 3D error distributions are presented for whole-body, body, hand and face in mm. We observe that 3D errors are mostly concentrated between 10mm and 20mm.

To assess the quality of the H3WB dataset, we conducted a cross-check study on 600 randomly selected images from the dataset. In this study, annotators were presented an image with the 2D projection of the 3D skeleton on top and were asked to manually correct mis-aligned keypoints by drag and drop. Using multi-view geometry, we reconstructed these corrected skeletons in 3D and compared them to our original skeletons. To validate our process, we show

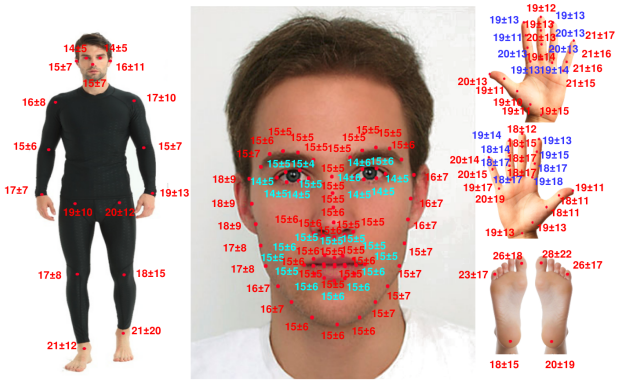

Figure 10. Per keypoints error statistics. Please zoom in.

the influence of each step in Table 2. The geometric approach produced good results but unfortunately cannot provide a large enough dataset. The completion step allows to obtain all labels but at the cost of degraded accuracy due to lack of alignment as explained in section 3.2. The diffusion recovers the original accuracy of the geometric approach. 2mm difference is irrelevant given the initial resolution of the images. We obtain a final **average error of 17mm** which is very accurate for such a difficult task, and leads to a benchmark which we believe will not be saturated until methods reach around 35mm.

Steps	# keypoints available	3D error (mm)			
		All	Body	Face	Hand
Geometry	48127	14.87	17.72	13.29	15.87
+ Completion	79800	29.31	25.57	26.02	36.67
+ Diffusion	79800	16.98	18.63	15.08	19.16

Table 2. Quantitative analysis of each intermediate step in our pipeline.

3D error distributions: In Figure 9, we illustrate 3D

error distributions in mm for all whole-body keypoints and each whole-body part keypoints (i.e. body, hand and face) separately. The distributions of errors are well concentrated around low values.

Per keypoint errors: All per keypoint error statistics are shown Figure 10. 98.3% of the images are below 5cm error (97.4% for body joints, 99.7% hands, and 96.4% face). Similarly, 82.6% of the images are below 2cm error (75.5% for body joints, 89.2% hands, and 75.8% face).

3.5. TotalCapture 3D WholeBody Dataset

In addition to H3WB dataset, we also prepare whole-body annotations for training sequence of poses from-TotalCapture [43] dataset using our proposed multi-view pipeline. We call this dataset TotalCapture 3D WholeBody, or T3WB for short. To create TotalCapture 3D WholeBody, we first obtain 2D whole-body keypoints from OpenPif-Paf [47]. At this stage we discard the frames without any human. Next, we finetune the completion network initialized by H3WB weights using 7000 samples of TotalCapture with complete 8-views. At the end we obtain 125,960 triplets of 3D whole-body keypoint, corresponding image and 2D projected keypoints. This shows our pipeline can be used with any multiview dataset.

We do not conduct a quality assessment study for T3WB and therefore we cannot guarantee the precision of the annotations. Instead, we use T3WB together with H3WB for training models. To alleviate the effect of the noisy annotations in T3WB, we sample more from H3WB than T3WB in each batch. More specifically, we follow 4:1 ratio for each batch during *H3WB + T3WB* trainings.

4. The H3WB benchmark

We use the H3WB dataset to propose a benchmark and the associated leaderboard. We split the dataset into training and test sets. The training set contains all samples from S1, S5, S6 and S7, including 80k {image,2D,3D} triplets. The test set contains all samples from S8, including 20k triplets. The test set labels are retained to prevent involuntary overfitting on the test set. Evaluation is accessible only by submitting results to the maintainers. We do not provide a validation set. We encourage researchers to report 5-fold cross-validation average and standard deviation (see supplementary).

The corresponding benchmark has 3 different tasks:

1. 3D whole-body lifting from complete 2D whole-body skeletons, or *2D→3D* for short.
2. 3D whole-body lifting from incomplete 2D whole-body skeletons, or *I2D→3D* for short.
3. 3D whole-body skeleton prediction from image, or *RGB→3D* for short.

For each task, we report the following MPJPE (Mean Per Joint Position Error) metrics:

- MPJPE for the whole-body, the body (keypoint 1-23), the face (keypoint 24-91) and the hands (keypoint 92-133) when whole-body is centered on the root joint, i.e. aligned with the pelvis, which in our case is the middle of two hip joints[2],
- MPJPE for the face when it is centered on the nose, i.e. aligned with keypoint 1,
- MPJPE for the hands when hands are centered on the wrist, i.e left hand aligned with keypoint 92 and right hand aligned with keypoint 113.

To create baselines on each task, we adapt popular methods from the literature by changing the number of keypoints to that of our whole-body dataset. Notice that we keep the training recipes of the original papers to avoid over-fitting to this new benchmark. In practice, we recommend to perform model selection and hyper-parameters tuning using 5-fold cross-validation.

4.1. 3D whole-body lifting from complete 2D whole-body keypoints (*2D→3D*)

This task is similar to the standard 3D human pose estimation from 2D keypoints but using whole-body keypoints. The training set contains 80k 2D-3D pairs. The test set contains only a half of all the test samples, *i.e*. 10k 2D poses[3].

We evaluate 6 methods on this task. SimpleBaseline [53] is a well-established model, consisting of a 6-layer MLP. We propose a modification, replacing the network architecture with an 8-layer MLP, which we call *Large Simple-Baseline* inspired by CanonPose [72]. Jointformer [52] is a recent transformer-based method. CanonPose is trained only with 2D supervision [72]. We also adapt CanonPose to work with additional 3D supervision by manually creating 3 fixed camera views and rotating the 3D skeletons into the corresponding view before projecting them into 2D, training it with multi-view weak-supervision. Finally, we report results for the parametric model SMPLify-X [60] by running optimizations on each input sample.

We train SimpleBaseline models using their official training setting as described in [53]. The inputs and targets are normalized by subtracting the mean and dividing by the standard deviation. Similarly, we train CanonPose [72] models following their official training setup where the inputs and targets are centered on the pelvis and scaled by the Forbenius norm. We train the Jointformer model in the two stages as described in [52].

[2]We provide the whole-body keypoint ids in supplementary material.

[3]The other half is reserved for the task I2D→3D to prevent access to the missing keypoints.

Method	All	Body	Face / aligned†	Hand / aligned‡
H3WB				
SMPL-X[60]	188.9	166.0	208.3 / 23.7	170.2 / 44.4
CanonPose[72]*	186.7	193.7	188.4 / 24.6	180.2 / 48.9
SimpleBaseline [53]*	125.4	125.7	115.9 / 24.6	140.7 / 42.5
CanonPose[72] *w* 3D sv.*	117.7	117.5	112.0 / 17.9	126.9 / 38.3
Large SimpleBaseline[53]*	112.3	112.6	110.6 / **14.6**	**114.8** / **31.7**
Jointformer[52]	**88.3**	**84.9**	**66.5** / 17.8	125.3 / 43.7
H3WB + T3WB				
CanonPose[72]*	164.7	161.1	174.5 / 21.5	150.8 / 43.6
SimpleBaseline [53]*	115.3	114.8	109.4 / **15.8**	125.1 / **33.5**
Jointformer[52]	**81.5**	**78.0**	**60.4** / 16.2	**117.6** / 38.8

Table 3. Comparing different methods for 2D→3D on H3WB test set. Results are shown for the MPJPE metric in mm. Methods with * output normalized predictions. Results of normalized methods are re-scaled using our scaling formula. All results are pelvis aligned, except † and ‡ show nose and wrist aligned results for face and hands, respectively. Sv. is supervision.

SimpleBaseline and CanonPose models output normalized whole-body keypoints which requires re-scaling at inference. We use statistics from the training set to adjust the test predictions. We calculate a scaling factor using the ratio of 3D to 2D bounding boxes. The formula is: $X_{\text{final}} = X_{\text{unit}} \times \overline{\sigma_{3d}} \times \frac{\sigma_{2d}}{\overline{\sigma_{2d}}}$, where X_{unit} is the normalized prediction, $\overline{\sigma_{3d}}$ is the average size of the 3D training boxes, σ_{2d} is the size of the current 2D box, and $\overline{\sigma_{2d}}$ is the average size of the 2D training boxes.

Since SMPLify-X has 144 keypoints with a different layout, we use interpolation to transform between the Whole-Body skeleton and SMPL-X and run SMPL-X's optimization for 2,000 iterations (4 minutes/sample).

We present the results in Table 3. SMPLify-X performs the worst, showing that parametric models struggle more than discriminative approaches. SimpleBaseline[53] is a solid method, and Large SimpleBaseline improves its performance further. CanonPose[72] can be improved with additional 3D supervision, but still performs worse than Large SimpleBaseline. CanonPose also predicts the camera view, and the uncertainty in this prediction can lead to more error. Jointformer[52] achieves the best results among all methods, but still has room for improvement. All methods perform worse on our benchmark than on Human3.6M because of pelvis centering, which creates higher numerical error on extremities like hands and face, the parts that contain most of the whole-body keypoints.

Additionally, we conducted experiments with SimpleBaseline, CanonPose, and Jointformer, leveraging the merged dataset *H3WB + T3WB*, which combines both H3WB and T3WB. The results in Table 3 show that when T3WB is integrated with H3WB, the performances are improved significantly. For instance, on all whole-body keypoints, it yields 22 pt, 10.1 pt and 6.8 pt improvement for SimpleBaseline, CanonPose and Jointformer, respectively.

4.2. 3D whole-body lifting from incomplete 2D whole-body keypoints (*I2D→3D*)

We propose a second task where we want to obtain 3D complete whole-body poses from 2D incomplete pose. This task aims to simulate the more realistic case when there are occlusions and the 2D whole-body detector outputs an incomplete skeleton. We do not provide masks for the training skeletons to allow for online data-augmentation. Instead, we propose a masking strategy as follows:

- With 40% probability, each keypoint has a 25% chance of being masked,
- with 20% probability, the face is entirely masked,
- with 20% probability, the left hand is entirely masked,
- with 20% probability, the right hand is entirely masked.

The second half of the test set (10k 2D) is devoted to this task. The masking strategy is applied only once on the 2D poses of the test set, which are directly provided as incomplete 2D skeletons for fair comparison between methods.

The results for the I2D→3D task are shown in Table 4. All methods perform worse than in the 2D→3D task. SimpleBaseline[53] has low capacity and uses batch normalization that struggles with missing data, resulting in poor performance. The Large SimpleBaseline model, without batch normalization layers, achieves good results for the task's complexity. CanonPose[72] performs poorly due to errors in camera rotation prediction, which are magnified since most of the 133 keypoints are on the face and hands. The addition of 3D supervision partly solves this problem. The transformer-based Jointformer[52] method outperforms others. Sample outputs obtained by Large SimpleBaseline are shown in Figure 11, where predicted skeletons, although not accurate, are realistic.

Similar to 2D→3D task, we experimented with SimpleBaseline, CanonPose and Jointformer on *H3WB + T3WB* dataset and obtained significant improvements. For all

Method	All	Body	Face / aligned†	Hand / aligned‡
CanonPose[72]*	285.0	264.4	319.7 / 31.9	240.0 / 56.2
SimpleBaseline[53]*	268.8	252.0	227.9 / 34.0	344.3 / 83.4
CanonPose[72] + 3D sv.*	163.6	155.9	161.3 / 22.2	171.4 / 47.4
Large SimpleBaseline[53]*	131.4	131.6	120.6 / **19.8**	**148.8** / **44.8**
Jointformer[52]	**109.2**	**103.0**	**82.4** / **19.8**	155.9 / 53.5
H3WB + T3WB				
CanonPose[72]*	261.5	243.3	291.3 / 31.3	223.1 / 53.7
SimpleBaseline[53]*	260.5	238.0	221.1 / 32.2	336.5 / 80.4
Jointformer[52]	**84.2**	**80.1**	**59.4** / **16.3**	**126.5**/ **44.5**

Table 4. Comparing different methods for I2D→3D on H3WB test set. Results are shown for the MPJPE metric in mm. Methods with * output normalized predictions. Results of normalized methods are re-scaled using our scaling formula. All results are pelvis aligned, except † and ‡ show nose and wrist aligned results for face and hands, respectively. Sv. is supervision.

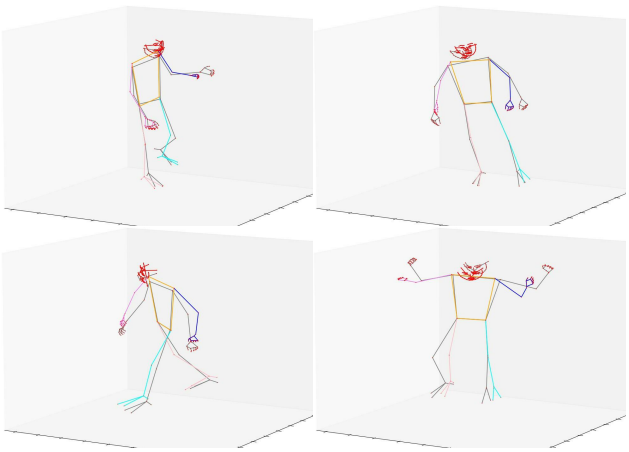

Figure 11. Example predictions from Large SimpleBaseline model for task I2D→3D. Colored skeletons correspond to predictions and gray skeletons correspond to groundtruths.

whole-body keypoints, the combined dataset yields 23.5 pt, 8.3 pt and 25 pt improvement for SimpleBaseline, CanonPose and Jointformer, respectively. Those results further validate the effectiveness of our multi-view pipeline for creating whole-body keypoint annotations for any multi-view dataset.

4.3. 3D whole-body pose estimation from a single image (*RGB→3D*)

This task is the standard monocular 3D human pose estimation task extended to whole-body pose estimation. We provide a script to split the original Human3.6M videos into images with our indexing in order to establish image-3D correspondences. The training set contains 80k {image paths,3D} pairs, as well as the 2D bounding box of the human in the image. The test set contains all the test samples, including 20k image paths and their 2D bounding boxes. 2D coordinates are not given in order to avoid collisions with 2D→3D and I2D→3D.

For this task, we train 2 two-stage models and 1 single-stage model. For our first two-stage model, we first train a Stacked Hourglass Network (SHN)[59] to predict 2D whole-body keypoints. Then, SimpleBaseline[53] takes 2D keypoint predictions as input and lifts them to 3D space. Similarly, the second two-stage model utilizes CPN[16] to output 2D keypoints and then Jointformer[52] lifts the 2D predictions to obtain 3D whole-body poses. For our single-stage model, we modify the last layer of Resnet50[33] to directly output the 3D whole-body keypoints. We regress the 3D whole-body keypoint coordinates using L1 loss.

Results in Table 5 show the two-stage *CPN + Jointformer* model obtains the best results. Our simple single-stage method performs better than the two-stage *SHN + SimpleBaseline* model. Learning 2D whole-body keypoints is challenging for SHN as very close keypoints on face and hands may introduce noise to the predicted keypoint heatmaps. The error in the 2D keypoints then makes the lifting task much more challenging. Surprisingly, RGB→3D seems to be harder that the I2D→3D task. Although there are also missing body parts due to self occlusion, RGB→3D contains more contextual information that should allow to better disambiguate the pose. Compared to 2D→3D and I2D→3D, direct prediction of 3D whole-body pose from images remains thus as a challenging task which we hope this benchmark can help improve over time.

Method	All	Body	Face / aligned †	Hand / aligned‡
RGB→2D+2D→3D:				
SHN[59]+SimpleBaseline*	182.5	189.6	138.7 / 32.5	249.4 / 64.3
CPN[16]+Jointformer[52]	**132.6**	**142.8**	**91.9 / 20.7**	**192.7 / 56.9**
RGB→3D:				
Resnet50[33]	166.7	151.6	123.6 / 26.3	244.9 / 63.1
DOPE[75]	191.3	199.7	187.3 / 66.0	193.3 / 78.2

Table 5. Comparing different methods for RGB→3D on H3WB test set. Results are shown for the MPJPE metric in mm. Methods with * output normalized predictions. Results of normalized methods are re-scaled using our scaling formula. All results are pelvis aligned, except † and ‡ show nose and wrist aligned results for face and hands, respectively.

In order to show the importance of training body parts jointly, we evaluate DOPE [75] on our benchmark. Unfortunately, it fails to address occluded body parts only predicts the whole-body keypoints for 35% of the test set. For each missing keypoint, we use the (topological) nearest predicted joint as a proxy. Even so, a disjointed model like DOPE fails to achieve significant accuracy.

5. Conclusion

In this paper, we introduce the H3WB dataset, which extends the Human3.6M dataset with 2D and 3D keypoint annotations for body, face, and hands, containing 100k images with 133 keypoints with an average accuracy of 17mm. We propose three tasks based on this dataset: 3D whole-body lifting from complete 2D keypoints, 3D whole-body lifting from incomplete 2D keypoints, and 3D whole-body prediction from monocular images. We evaluate several baselines on these tasks and demonstrate promising accuracy, but with room for improvement. Lifting from incomplete 2D skeletons and direct estimation from monocular images remain challenging, and we hope that our dataset and benchmark will spur future research in these areas.

Acknowledgments

This work was supported by ANR project TOSAI ANR-20-IADJ- 0009 and Ergonova Conseil, and was granted access to the HPC resources of IDRIS under the allocation 2023-AD011012640R2 and 2023-AD011013267R1 made by GENCI.

References

[1] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 2019. 1

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *CVPR*, 2019. 2

[3] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *TOG*, 2021. 1

[4] Dennis Bautembach, Iason Oikonomidis, and Antonis Argyros. Filling the joints: Completion and recovery of incomplete 3d human poses. *Technologies*, 2018. 3

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999. 3

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. *ECCV*, 2016. 2

[7] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, 2019. 1, 2

[8] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, 2018. 1, 2

[9] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In *CVPR*, 2018. 1

[10] Nicolò Carissimi, Paolo Rota, Cigdem Beyan, and Vittorio Murino. Filling the gaps: Predicting missing joints of human poses using denoising autoencoders. In *ECCV Workshops*, 2018. 3

[11] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016. 5

[12] Cristian Sminchisescu Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. In *ICCV*, 2011. 2

[13] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Expnet: Landmark-free, deep, 3d facial expressions. In *FG 2018*, 2018. 1, 3

[14] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *CVPR*, 2021. 2

[15] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *CVPR*, 2021. 2

[16] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *CVPR*, 2017. 9

[17] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, 2021. 2

[18] 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, 3

[19] Daniel Crispell and Maxim Bazik. Pix2face: Direct 3d face model estimation. In *ICCV Workshops*, pages 2512–2518, 2017. 3

[20] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. *CVPR Workshop*, 2019. 3

[21] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, pages 534–551, 2018. 3

[22] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 1

[23] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1343–1351, 2021. 1

[24] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9919–9928, 2021. 1

[25] Mercedes Garcia-Salguero, Javier Gonzalez-Jimenez, and Francisco-Angel Moreno. Human 3d pose estimation with a tilting camera for social mobile robot interaction. *Sensors*, 2019. 1

[26] Erik Gärtner, Aleksis Pirinen, and Cristian Sminchisescu. Deep reinforcement learning for active human pose estimation. *AAAI*, 2020. 1

[27] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, pages 10833–10842, 2019. 1, 2

[28] Yiwen Gu, Shreya Pandit, Elham Saraee, Timothy Nordahl, Terry Ellis, and Margrit Betke. Home-based physical therapy with an interactive computer vision system. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 1

[29] Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José MF Moura, and Manuela Veloso. Teaching robots to predict human motion. In *IROS*, 2018. 1

[30] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose

estimation using explicit 2d features and intermediate 3d representations. *CVPR*, 2019. 1

[31] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Handsformer: Keypoint transformer for monocular 3d pose estimation ofhands and object in interaction. *CVPR*, 2022. 1

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 4

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2015. 9

[34] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. *CVPR*, 2020. 2

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 5

[36] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPMAI*, 2014. 2, 3

[37] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, 2018. 1

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

[39] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *ICCV*, 2017. 3

[40] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 2, 3

[41] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2

[42] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 2

[43] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. *CVPR*, 2018. 2, 3, 7

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

[45] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. *ICCV*, 2019. 1

[46] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 1

[47] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 3, 7

[48] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 2

[49] 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. 2

[50] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *CVPR*, 2020. 2

[51] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015. 2

[52] Sebastian Lutz, Richard Blythman, Koustav Ghosal, Matthew Moynihan, Ciaran Simms, and Aljosa Smolic. Jointformer: Single-frame lifting transformer with error prediction and refinement for 3d human pose estimation. *ArXiv*, 2022. 7, 8, 9

[53] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 2, 7, 8, 9

[54] 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. 2

[55] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018. 1, 2

[56] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single RGB image. *ICCV*, 2019. 1, 2

[57] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single RGB image. *ECCV*, 2020. 2

[58] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, 2018. 1, 2

[59] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *ECCV*, 2016. 9

[60] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed Osman, Dimitrios Tzionas, and Michael Black. Expressive body capture: 3D hands, face, and body from a single image. *CVPR*, 2019. 1, 3, 7, 8

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

[62] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[63] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *TPAMI*, 2019. 1, 2

[64] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2

[65] Yu Rong, Ziwei Liu, and Chen Change Loy. Chasing the tail in monocular 3d human reconstruction with prototype memory. *TIP*, 2022. 1

[66] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. *ICCV*, 2021. 1, 3

[67] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 1

[68] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *CVPR*, 2019. 1, 3

[69] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, 2017. 3

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017. 4

[71] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2

[72] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *CVPR*, 2021. 2, 7, 8

[73] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *CVPR*, 2021. 1

[74] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *ECCV*, 2020. 1

[75] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. DOPE: distillation of part experts for whole-body 3d pose estimation in the wild. *ECCV*, 2020. 1, 3, 9

[76] Yandong Wen, Weiyang Liu, Bhiksha Raj, and Rita Singh. Self-supervised 3d face reconstruction via conditional estimation. *ICCV*, 2021. 3

[77] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 1, 2, 3

[78] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *CVPR*, 2020. 2

[79] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. Aligning latent spaces for 3d hand pose estimation. In *ICCV*, 2019. 2

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

[81] Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human synthesis and scene compositing. *AAAI*, 2019. 1

[82] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, pages 2354–2364, 2019. 1, 2

[83] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *CVPR*, 2021. 3

[84] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single RGB images. *ICCV*, 2017. 1, 2

[85] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan C. Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single RGB images. *ICCV*, 2019. 2