Predicting Camera Viewpoint Improves Cross-dataset Generalization for 3D Human Pose Estimation

Zhe Wang, Daeyun Shin, and Charless C. Fowlkes

University of California, Irvine
{zwang15,daeyuns, fowlkes}@ics.uci.edu
http://wangzheallen.github.io/cross-dataset-generalization

Abstract. Monocular estimation of 3d human pose has attracted increased attention with the availability of large ground-truth motion capture datasets. However, the diversity of training data available is limited and it is not clear to what extent methods generalize outside the specific datasets they are trained on. In this work we carry out a systematic study of the diversity and biases present in specific datasets and its effect on cross-dataset generalization across a compendium of 5 pose datasets. We specifically focus on systematic differences in the distribution of camera viewpoints relative to a body-centered coordinate frame. Based on this observation, we propose an auxiliary task of predicting the camera viewpoint in addition to pose. We find that models trained to jointly predict viewpoint and pose systematically show significantly improved cross-dataset generalization.

Keywords: monocular 3d human pose estimation, cross dataset evaluation, dataset bias.

1 Introduction

A large swath of computer vision research increasingly operates in playing field which is swayed by the quantity and quality of annotated training data available for a particular task. How well do you know your data? Fig 1 presents a sampling images from 5 popular datasets used for training models for 3d human pose estimation (Human3.6M [8], GPA [43], SURREAL [40], 3DPW [15] , 3DHP [19]). We ask the reader to consider the game of "Name That Dataset" in homage to Torralba *et al.* [33]. Can you guess which dataset each image belongs to? More importantly, if we train a model on the Human3.6M dataset (at Fig 1 left) how well would you expect it to perform on each of the images depicted?

Each of these datasets were collected using different mocap systems (VICON, The Capture, IMU), different cameras (Kinect, commercial synchronized cameras, phone), and collected in different environments (controlled lab environment, marker-less in the wild environment, or synthetic images) with varying camera viewpoint and pose distributions (see Fig 3). These datasets contain further variations in body sizes, camera intrinsic and extrinsic parameters, body and

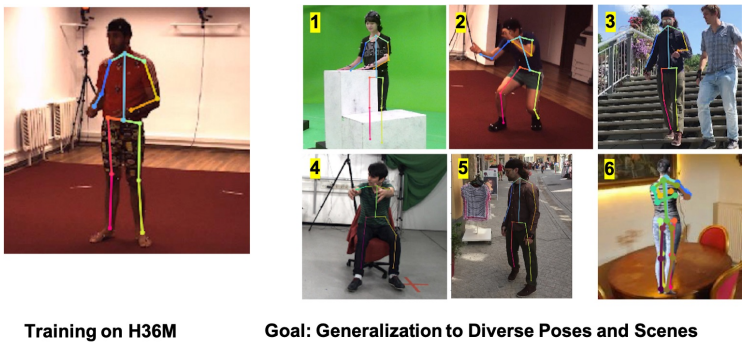


Fig. 1: In this paper we consider the problem of dataset bias and cross-dataset generalization. Can you guess which human pose dataset each image on the right comes from? If we train a model on H36M data (left) can you predict which image has the lowest/highest 3D pose prediction error? (answer key below)[1]

background appearance. Despite the obvious presence of such systematic differences, these variables and their subsequent effect on performance have yet to be carefully analyzed.

In this paper, we study the generalization of 3d pose models across multiple datasets and propose an auxiliary prediction task: estimating the relative rotation between camera viewing direction and a body-centered coordinate system defined by the orientation of the torso. This task serves to significantly improve cross-dataset generalization. Ground-truth for our proposed camera viewpoint task can be derived for existing 3D pose datasets without requiring additional labels. We train off-the shelf models [21, 50] which estimate the camera-relative 3d pose, augmented with a viewpoint prediction branch. In our experiments, we show our approach outperforms the state-of-the-art PoseNet [21] and [50] baseline by a large margin across 5 different 3d pose datasets. Perhaps even more startling is that the addition of this auxiliary task results in significant improvement in cross-dataset test performance. This simple approach increases robustness of the model and, to our knowledge, is the first work that systematically confronts the problem of dataset bias in 3d human pose estimation.

To summarize, our main contributions are:

• We analyze the differences among contemporary 3d human pose estimation datasets and characterize the distribution and diversity of viewpoint and body-centered pose.

• We propose the novel use of camera viewpoint prediction as an auxiliary task that systematically improves model generalization by limiting overfitting

[1] Answer key: Metric: MPJPE, the lower the better. 1) GPA: 69.7 mm 2) H36M: 29.2 mm, 3) 3DPW, 71.2 mm, 4) 3DHP 107.7 mm, 5) 3DPW 66.2 mm, 6) SURREAL 83.4 mm, H36M image performs best while 3DHP image performs worst.

to common viewpoints and can be directly calculated from commonly available joint coordinate ground-truth.

• We experimentally demonstrate the effectiveness of the viewpoint prediction branch in improving cross-dataset 3d human pose estimation over two popular baseline and achieve state-of-the-art performance on five datasets.

2 Related Work

Cross-Dataset Generalization and Evaluation 3d human pose estimation from monocular imagery has attracted significant attention due to its potential utility in applications such as motion retargeting [41], gaming, sports analysis, and health care [16]. Recent methods are typically based on deep neural network architectures [17, 21, 23, 24, 30, 50] trained on one of a few large scale, publicly available datasets. Among these are [17, 23, 30] evaluated on H36M, [19, 50] work on both H36M [8] and 3DHP [19], [15, 34] work on TOTALCAPTURE [34] and 3DPW[15], [43] work on the GPA dataset [43]. [40] works on both SURREAL [40] and H36M [8] dataset.

Given the powerful capabilities of CNNs to overfit to specific data, we are inspired to revisit the work of [33], which presented a comparative study of popular object recognition datasets with the goals of improving dataset collection and evaluation protocols. Recently, [13] observed characteristic biases present in commonly used depth estimation datasets and proposed scale invariant training objectives to enable mixing multiple, otherwise incompatible datasets. [52] introduced the first large-scale, multi-view unbiased hand pose dataset as training set to improve performance when testing on other dataset. Instead of proposing yet another dataset or resorting to domain adaptation approaches (see e.g., [42]), we focus on identifying systematic biases in existing data and identifying generic methods to prevent overfitting in 3d pose estimation.

Coordinate Frames for 3D Human Pose In typical datasets, gold-standard 3d pose is collected with motion capture systems [8, 29, 34, 43] and used to define ground-truth 3D pose relative one or more calibrated RGB camera coordinate systems [8, 15, 19, 40, 43]. To generate regression targets for use in training and evaluation, it is typical to predict the *relative* 3d pose and express the joint positions relative to a specified root joint such as the pelvis (see e.g.,[21, 30]). We argue that camera viewpoint is an important component of the experimental design which is often overlooked and explore using a body-centered coordinate system which is rotated relative to the camera frame.

This notion of view-point invariant prediction has been explored in the context of 3D object shape estimation [3, 4, 20, 26, 28, 31, 36, 46] where many works have predicted shape in either an object-centered or camera-centered coordinate frame [28, 32, 49]. Closer to our task is the 3d hand pose estimator of [51] which separately estimated the viewpoint and pose (in canonical hand-centered coordinates similar to ours) and then combine the two to yield the final pose in the camera coordinate frame. However, we note that predicting canonical

pose directly from image features is difficult for highly articulated objects (indeed subsequent work on hand pose, e.g. [38], abandoned the canonical frame approach). Our use of body-centered coordinate frames differs in that we only use them as a auxiliary training task that improves prediction of camera-centered pose.

3D Human Pose Estimation With the recent development of deep neural networks (CNNs), there are significant improvements on 3D human pose estimation [5, 17, 23, 44]. Many of them try to tackle in-the-wild images. [50] proposes to add bone length constraint to generalize their methods to in the wild image. [27] seeks to pose anchors as classification template and refine the prediction with further regression loss. [5] propose a a new disentangled hidden space encoding of explicit 2D and 3D features for monocular 3D human pose estimation that shows high accuracy and generalizes well to in-the-wild scenes, however, they do not evaluate its capacity on indoor cross-dataset generalization. To the best of our knowledge, our work is the first to exploit cross-dataset task not only towards in-the-wild generalization but also across different indoor datasets.

Multi-task Training There have has been a wide variety of work in training deep CNNs to perform multiple tasks, for example: joint detection, classification, and segmentation [6], joint surface normal, depth, and semantic segmentation [12], joint face detection, keypoint, head orientation and attributes [25]. Such work typically focuses on the benefits (accuracy and computation) of jointly training a single model for two or more related tasks. For example, predicting face viewpoint has been shown to improve face recognition [47]. Our approach to improving generalization differs in that we train models to perform two tasks (viewpoint and body pose) but discard viewpoint predictions at test time and only utilize pose. In this sense our model is more closely related to work on "deeply-supervised" nets [14, 45] which trains using losses associated with auxiliary branches that are not used at test time.

3 Variation in 3D Human Pose Datasets

We begin with a systematic study of the differences and biases across 3d pose datasets. We selected three well established datasets Human3.6m (H36M), MPI-inf-3dhp (3DHP), SURREAL, as well as two more recent datasets 3DPW and GPA for analysis. These are large-scale datasets with a wide variety of characteristics in terms of capture technology, appearance (in-the-wild,in-the-lab,synthetic) and content (range of body sizes, poses, viewpoints, clothing, occlusion and human-scene interaction). In this paper, we focus on characterizing variation in geometric quantities (pose and viewpoint) which can be readily quantified (compared to, e.g., lighting and clothing).

We list some essential statistics from 5 datasets in Table 1. For these datasets, gold-standard 3d pose is collected with motion capture systems [8, 29, 34, 43] and used to define ground-truth 3D pose relative one or more calibrated RGB camera coordinate systems [8, 15, 19, 40, 43]. To generate regression targets for use in

Dataset	H36M	GPA	SURREAL	3DPW	3DHP
Year	2014	2019	2017	2018	2017
Imaging Space	1000 × 1002	1920 × 1080	320 × 240	1920 × 1080	2048 × 2048 or 1920 × 1080
Camera Distance	5.2 ± 0.8	5.1 ± 1.2	8.0 ± 1.0	3.5 ± 0.7	3.8 ± 0.8
Camera Height	1.6 ± 0.05	1.0 ± 0.3	0.9 ± 0.1	0.6 ± 0.8	0.8 ± 0.4
Focal Length	1146.8 ± 2.0	1172.4 ± 121.3	600 ± 0	1962.2 ± 1.5	1497.88 ± 2.8
No. of Joints	38	34	24	24	28 or 17
No. of Cameras	4	5	1	1	14
No. of Subjects	11	13	145	18	8
Bone Length	3.9 ± 0.1	3.7 ± 0.2	3.7 ± 0.2	3.7 ± 0.1	3.7 ± 0.1
GT source	VICON	VICON	Rendering	SMPL	The Capture
No. Train Images	311,951	222,514	867,140	22,375	366,997
No. Test Images	109,764	82,378	507	35,515	2,875

Table 1: Comparison of existing datasets commonly used for training and evaluating 3D human pose estimation methods. We calculate the mean and std of camera distance, camera height, focal length, bone length from training set. Focal length is in mm while the others are in unit meters. 3DHP has two kinds of cameras and the training set provide 28 joints annotation while test set provide 17 joints annotation.

training and evaluation, it is typical to predict the *relative* 3d pose (see e.g.,[21, 30]) and express the joint positions relative to a specified root joint (typically the pelvis) and crop/scale the input image accordingly. This pre-processing serves to largely "normalize away" dataset differences in camera intrinsic parameters and camera distance shown in Table 1. However, it does not address camera orientation.

To characterize the remaining variability, we factor the camera-relative pose into camera viewpoint (the position of the camera relative to a canonical body-centered coordinate frame defined by the orientation of the person's torso) and the pose relative to this body-centered coordinate frame.

Computing Body-centered Coordinate Frames To define a viewpoint-independent pose, we need to specify a canonical body-centered coordinate frame. As shown in Fig 9a, we take the origin to be the camera-centered coordinates of root joint (pelvis) $p_p = (x_p, y_p, z_p)$ and the orientation is defined by the plane spanned by p_p, the left shoulder p_l and the right shoulder p_r. Given these joint positions, we can compute an orthogonal frame consist-

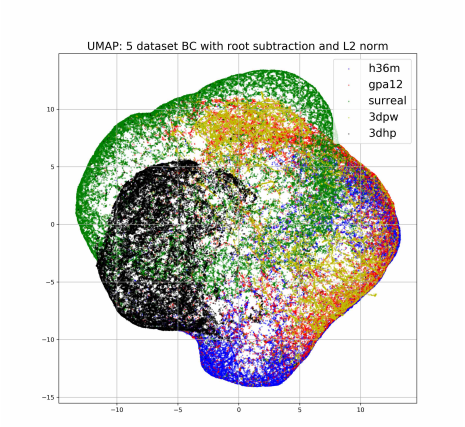
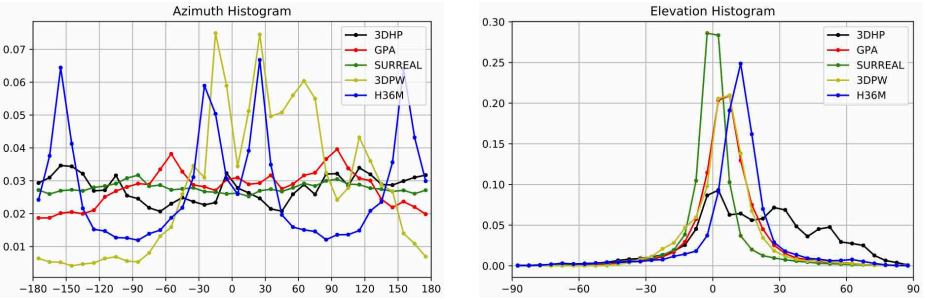


Fig. 2: Distribution of view-independent body-centered pose, visualized as a 2D embedding produced with UMAP [18]



(a) Viewpoint Azimuth (b) Viewpoint Elevation

Fig. 3: Distribution of camera viewpoints relative to the human subject. We show the distribution of camera azimuth $(-180°, 180°)$ and elevation $(-90°, 90°)$ for 50k poses sampled from each representative dataset (**H36M**, **GPA**, **SURREAL**, **3DPW**, **3DHP**).

ing of the front direction f, up direction u and right direction r are defined as:

$$u = (p_l + p_r)/2 - p_p$$
$$f = (p_l - p_p) \times (p_r - p_p)$$
$$r = f \times u$$

The rotation between the body-centered frame and the camera frame is then given by the matrix $R = -[r, u, f]$. We find it useful to represent rotations using unit quaternions (as have others, e.g. [41, 35]). The corresponding unit quaternion representing R has components:

$$q = \frac{1}{4q_0}[4q_0^2, u_2 - f_1, f_0 - r_2, r_1 - u_0], \qquad q_0 = \sqrt{(1 - r_0 - u_1 - f_2)} \tag{1}$$

Distribution of Camera Viewpoints Fig 3 shows histograms capturing the distribution of camera viewing direction in terms of azimuth (Fig 3a) and elevation (Fig 3b) relative to the body-centered coordinate system for 50k sample poses from each of the 5 datasets.

We observe **H36M** has a wide range of view direction over azimuth with four distinct peaks (-30 degree, 30 degree, -160 degree, 160 degree), it shows during the capture session subjects are always facing towards or facing away the control center while the four RGB cameras captured from four corners. H36M has a clear bias towards elevation above 0; **GPA** is more spread over azimuth compared with H36M, most of the views range from -60 degree to 90 degree; **SURREAL** synthetically sampled camera positions with a uniform distribution over azimuth, and also have a uniform distribution over elevation. The viewpoint bias for **3DPW** arises naturally from filming people in-the-wild from a handheld or tripod mounted camera roughly the same height as the subject. Of the non-synthetic datasets, **3DHP** is the most uniform spread over azimuth and includes

a wider range of positive elevations, a result of utilizing cameras mounted at multiple heights including the ceiling.

These differences are further highlighted in Fig 9 which shows the joint distribution of camera views and reveals the source of non-uniformity of the azmuthal distribution for 3DHP and H36M due to subjects tending to face a canonical direction while performing some actions. For example, in H36M in Fig 9b, actions in which the subject lean over or lie down (extreme elevations) only happen at particular azimuths. Similarly, in 3DHP (Fig 9f), the 14 camera locations are visible as dense clusters at specific azimuths indicating a significant subset of the data in which the subject was facing in a canonical direction relative to the camera constellation.

Distribution of Pose To characterize the remaining variability in pose after the viewpoint is factored out, we used the coordinates of 14 joints common to all datasets expressed in the body-centered coordinate frame. We also scaled the body-centered joint locations to a common skeleton size (removing variation in bone length shown in Table 1). To visualize the resulting high-dimensional data distribution, we utilized UMAP [18] to perform a non-linear embedding into 2D. Figure 2 shows the resulting distributions which show a substantial degree of overlap. For comparison, please see the Appendix which show embeddings of the same data when bone length and/or viewpoint are not factored out.

We also trained a multi-layer perceptron to predict which dataset a given body-relative pose came from. It had an average test accuracy of 20% providing further evidence of relatively little bias in the distribution of poses across datasets once viewpoint and body size are factored out.

4 Learning Pose and Viewpoint Prediction

To overcome biases in viewpoint across datasets, we propose to use viewpoint prediction as an auxiliary task to regularize the training of standard camera-centered pose estimation models.

4.1 Baseline architecture

Our baseline model [21, 50] consists of two parts: the first ResNet [7] backbone which takes in images patches cropped around the human; followed by the second part which takes the resulting feature map and upsamples it using three consecutive deconvolutional layers with batch normalization and ReLU. A 1-by-1 convolution is applied to the upsampled feature map to produce the 3D heatmaps for each joint location. The soft-argmax [30] operation is used to extract the 2D image coordinates $(\hat{x}_j, \hat{y}_j)$ of each joint j within the crop, and the root-relative depth $\hat{y}_j$. At test time, we can convert this prediction into into a 3d metric joint location $p_j = (x_j, y_j, z_j)$ using the crop bounding box, an estimate of the root joint depth or skeleton size, and the camera intrinsic parameters.

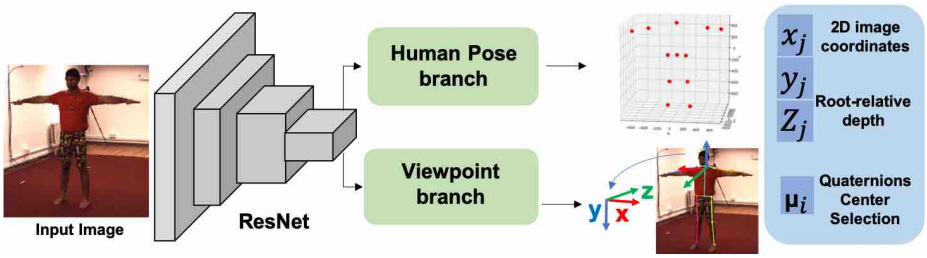


Fig. 4: Flowchart of our model. We augment a model which predicts camera-centered 3d pose using the **human pose branch** with an additional **viewpoint branch** that selections among a set of quantized camera view directions.

The loss function of the coordinate branch is the $L1$ distance between the estimated and groud-truth coordinates.

$$\mathcal{L}_{pose} = \frac{1}{J}\sum_{j=1}^{J} ||p_j - p_j^*||_1$$

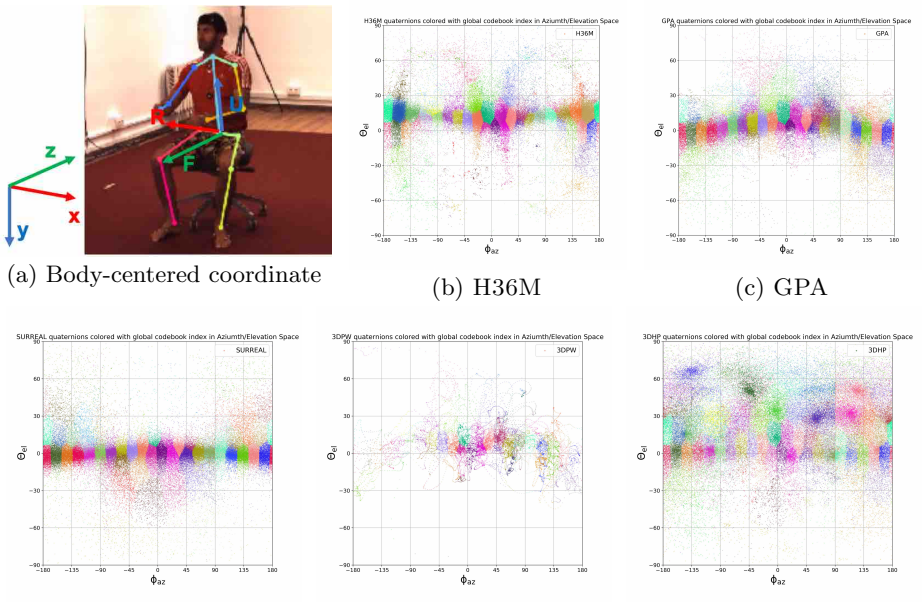
4.2 Predicting the camera viewpoint

To predict the camera viewpoint relative to the body-centered coordinate frame we considered three approaches: (i) direct regression of q, (ii) quantizing the space or rotations and performing k-way classification, and (iii) a combined approach of first predicting a quantized rotation followed by regressing the residual from the cluster center. In our experiments, we found that the classification-based loss yields less accurate coordinate frame predictions but yielded the largest improvements in the pose prediction branch (see Table 4).

To quantize the space of rotations, we use k-means to cluster the quaternions into k=100 clusters. The clusters are computed from training data of a single dataset (local clusters) or from all five datasets (global clusters). We visualize the global cluster centers in azimuth and elevation space in Fig 9 b-f, as well as randomly sampled quaternions from H36M, GPA, SURREAL, 3DPW and 3DHP datasets.

To regress the quaternion q we simply add a branch to our base pose prediction model consisting of a 1x1 convolutional layer to reduce the feature dimension to 4 followed by global average pooling and normalization to yield a unit 4-vector. We train this variant using a standard squared-Euclidan loss on target q^*. For classification, we use the same prediction q but compute the probability it belongs to the correct cluster using a softmax to get a distribution over cluster assignments:

$$p(c|q) = \frac{\exp(-\mu_c^T q)}{\sum_{i=1}^{k}\exp(-\mu_i^T q)}$$



(a) Body-centered coordinate (b) H36M (c) GPA

(d) SURREAL (e) 3DPW (f) 3DHP

Fig. 5: **a**: Illustration of our body-centered coordinate frame (up vector, right vector and front vector) relative to a camera-centered coordinate frame. **b-f**: Camera viewpoint distribution of the 5 datasets color by quaternion cluster index. Quaternions (rotation between body-centered and camera frame) are sampled from training sets and clustered using k-means. They are also visualized in azimuth / elevation space following Fig 3.

where $\{\mu_1, \mu_2, \ldots, \mu_k\}$ are the quaternions corresponding to cluster centers computed by k-means. We use the negative log-likelihood as the training loss,

$$\mathcal{L}_q = -log(p(c^*|q))$$

where c^* is the viewpoint bin that the training example was assigned during clustering. Our final loss consists of both quaternion and pose terms: $\mathcal{L} = \lambda\mathcal{L}_q + \mathcal{L}_{pose}$.

5 Experiments

Data and evaluation metric. To reduce the redundancy of the training images (30 fps video gives lots of duplicated images for network training), we down sample 3DHP, SURREAL to 5 fps. Following [21, 50], we sample H36M to 10 fps, and use the protocol 2 (subject 1,3,5,7,8 for training and subject 9,11 for testing) for evaluation. As GPA is designed as monocular image 3d human pose estimation, which is already sampled, we follow [43] and directly use the released set. Number of images in train set and test set is shown in Table 1. In addition, we use the MPII dataset [1], a large scale in-the-wild human pose dataset for training a more

		MPJPE (in mm, lower is better)				
	Testing \Training	H36M	GPA	SURREAL	3DPW	3DHP
Baseline	H36M	**53.2**	110.5	107.1	125.1	108.4
	GPA	105.2	**53.9**	86.8	111.7	90.5
	SURREAL	118.6	103.2	**37.2**	120.8	108.2
	3DPW	108.7	116.4	114.2	**100.6**	113.3
	3DHP	111.8	123.9	120.3	139.7	**91.9**
Our Method	H36M	**52.0**	102.5	103.3	124.2	95.6
	GPA	98.3	**53.3**	85.6	110.2	91.3
	SURREAL	114.0	101.2	**37.1**	113.8	107.2
	3DPW	109.5	112.0	112.2	**89.7**	105.9
	3DHP	111.9	119.7	118.2	136.0	**90.3**
Same-Dataset Error Reduction ↓		1.2	0.6	0.1	10.9	1.5
Cross-Dataset Error Reduction ↓		10.6	18.6	9.1	13.1	20.4

Table 2: Baseline cross-dataset test error and error reduction from the addition of our proposed quaternion loss. Bold indicates the best performing model on each the test set (rows). Blue color indicates test set which saw greatest error reduction. See appendix for corresponding tables of PCK and Procrustese aligned MPJPE.

robust pose model. It contains 25k training images and 2,957 validation images. We use two metrics, first is mean per joint position error (MPJPE), which is calculated between predicted pose and ground truth pose. The second one is PCK3D [19], which is the accuracy of joint prediction (threshold on MPJPE with 150mm).

Implementation Details. As different datasets have diverse joint configuration, we select a subset of 14 joints that all datasets share to eliminate the bias introduced by different number of joints during training. We normalize the z value from $(-z_{max}, +z_{max})$ to $(0, 63)$ for integral regression. z_{max} is 2400 mm based all 5 set. We use PyTorch to implement our network. The ResNet-50 [7] backbone is initialized using the pre-trained weights on the ImageNet dataset. We use the Adam [11] optimizer with a mini-batch size of 128. The initial learning rate is set to 1×10^{-3} and reduced by a factor of 10 at the 17th epoch, we train 25 epochs for each of the dataset. We use 256×256 as the size of the input image of our network. We perform data augmentation including rotation, horizontal flip, color jittering and synthetic occlusion following [21]. We set λ to 0.5 for the quaternion loss which is validated on 3DPW validation set.

5.1 Cross-dataset evaluation

We list the cross-dataset baseline and our improved results in Table 2. The bold numbers indicate the best performing model on the test set. As expected, the best performance occurs when the model is trained and evaluated on the same set. The numbers marked with blue color indicate the test set where the error reduction is most significant, using our proposed quaternion loss.

Training on H36M. Adding the quaternion loss reduces total cross-dataset error by 10.6 mm (MPJPE), while the same-dataset error reduction is 1.2 mm (MPJPE).

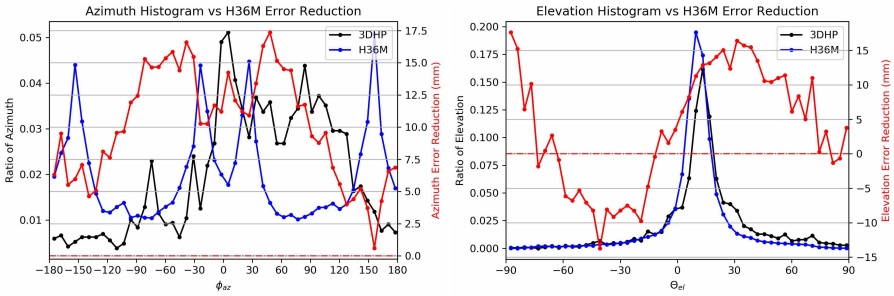


Fig. 6: We visualize viewpoint distributions for train (3DHP) and test (H36M) overlayed with the reduction in pose prediction error relative to baseline

	MPJPE (in mm, lower is better)				
Metric \Training Set	H36M	GPA	SURREAL	3DPW	3DHP
Same-Dataset Error Reduction ↓	0.6	4.2	0.2	7.6	1.2
Cross-Dataset Error Reduction ↓	2.4	12.3	1.9	10.1	9.3

Table 3: Retraining the model of Zhou *et al.* [50] using our viewpoint prediction loss yields also shows significant decrease in prediction error, demonstrating the generality of our finding. See appendix for full table of numerical results.

This may be explained by the error on H36M already being low. The largest error reduction is on GPA (6.9 mm) which we attribute to de-biasing the azimuth distribution difference as shown in Fig 3a.

Training on GPA. The total cross-dataset error reduction is 18.6 mm (MPJPE), and the same data error reduction is 0.6 mm (MPJPE). We attribute this to the bias during capture [43]: the coverage of camera viewing directions is centered in the range of −60 to 90 degrees azimuth (as in Fig 3a). The largest cross-data set error reduction occurs for H36M, with 8.0 mm. This further demonstrates that the view direction distribution is largely different from H36M.

Training on SURREAL. Adding the quaternion loss reduces the cross-dataset error by 9.1 mm (MPJPE), while the same-dataset error reduction is 0.1 mm (MPJPE). We attribute this to the fact that viewpoint distribution on SURREAL itself is already uniform as in Fig 3a. We can see distribution over azimuths is quite uniform. Thus adding more supervision in the form of quaternion loss helps little. The most error reduction (2.0mm) is observed on 3DPW. We attribute this to the fact that 3DPW is strongly biased dataset in terms of view direction, and the quaternion loss helps reduce the view difference between SURREAL and 3DPW.

Training on 3DPW. The error is reduced by 10.9 mm (MPJPE) on itself (also the most error reduction one with model trained on 3DPW), and the cross-dataset error reduction is 13.1 mm (MPJPE). From the Fig 3a we can see, in terms of

Datasets	Baseline	C	R	C+R	C+local cluster	C+cannonical pose
3DPW (MPJPE (mm))	100.6	89.7	94.0	93.2	93.1	100.3

Table 4: Ablation analysis: we compare the performance of our proposed camera view-point loss using classification (C), regression (R), using both (C+R); using per-dataset clusterings (local) rather than the global clustering; and adding a third branch which also predicts pose in canonical body-centered coordinates.

azimuth, 3DPW has a strong bias towards −30 degree to 60 degree. As during capture, the subject is always facing towards the camera to make it easier for association between the subject (there are multiply persons in crowded scene) and IMU sensors, this bias seems inevitable and quaternion loss is helpful for this kind of in the wild dataset to reduce view direction bias. It is also verified in 3DHP, where half of the test set is in the wild, and have view direction bias.

Training on 3DHP. Adding the quaternion loss reduces the total cross-dataset error by 20.4 mm, while the same-dataset error reduction is 1.5 mm (MPJPE). During the capture, 3DHP capture images from a wide range of viewpoints. We can see from the Fig 3 that the azimuth of 3DHP is the most uniformly distributed of the real datasets. Thus treating it as training set will enable the network to be robust to view direction. We also calculate error reduction conditioned on azimuth and elevation on the H36M test set (Fig 6). The blue/black line is azimuth and elevation histogram distribution for H36M/3DHP training sets while the red line shows relative error reduction for H36M. We can see the error is reduced more where H36M has fewer views relative to 3DHP.

5.2 Effect of Model Architecture and Loss Functions

To demonstrate the generalization of our approach to other models, we also added a viewpoint prediction branch to the model of [50] which utilizes a different model architecture. We observe similar results in terms of improved generalization (see Table 3 and appendix). We note that while our primary baseline model [21] uses camera intrinsic parameters to back-project, [50] utilizes an average bone-length estimate from the training set which results in higher prediction errors across datasets.

Ablation study To explore whether our methods are robust to different k-means initialization, we repeat k-means 4 times and report performance on 3DPW. We find the range of the MPJPE is within 90 ± 0.4 ([89.9, 89.6, 90.2, 89.7]) mm. We also vary the number of clusters to select the best $k \in \{10, 24, 50, 100, 200, 500\}$, with corresponding errors [93.0, 95.2, 92.3, 89.7, 93.0,93.2]. We find k=100 is the best number with at most 6 mm reduction compared to k=24. In Table 4, the error of global clusters is 3.4 mm error less than local, per-dataset clusters, demonstrating training on global clusters is better than local clusters which are biased towards the training set view distribution. In terms of choice for

	MPJPE↓: lower is better					PCK3D↑: higher is better				
	H36M	GPA	SURREAL	3DPW	3DHP	H36M	GPA	SURREAL	3DPW	3DHP
Mehta [19]	72.9	-	-	-	-	-	-	-	-	64.7
Zhou [50]	64.9	96.5	-	-	-	-	82.9	-	-	72.5
Arnab[2]	77.8	-	-	-	-	-	-	-	-	-
Kanazawa [9]	88.0	-	-	-	124.2	-	-	-	-	72.9
Kanazawa [10]	-	-	-	127.1	-	-	-	-	**86.4***	-
Moon [21]	54.3	-	-	-	-	-	-	-	-	-
Kolotouros [22]	78.0	-	-	-	-	-	-	-	-	-
Tung[37]	98.4	-	64.4*	-	-	-	-	-	-	-
Varol[39]	**51.6***	-	49.1	-	-	-	-	-	-	-
Habibie [5]	65.7	-	-	-	91.0	-	-	-	-	82.0
Yu [48]	59.1	-	-	-	-	-	-	-	-	-
Ours	52.0	**53.3**	**37.1**	**89.7**	**90.3**	**96.0**	**96.8**	**97.3**	84.6	**84.3**

Table 5: Comparison to state-of-the-art performance. There are many missing entries, indicating how infrequent it is to perform multi-dataset evaluation. Our model provides a new state-of-the art baseline across all 5 datasets and can serve as a reference for future work. * denotes training using extra data or annotations (e.g. segmentation). Underline denotes the second best results.

quaternion regression, k-way classification reduced error by 4.3 mm compared to regression. While utilizing both classification and regression losses gives error than regression only.

Finally, we also consider adding a third branch and loss function to the model which also predicts the 3D pose in the body-centered coordinate system. This is related to the hand pose model of [51], although we don't use this prediction of canonical pose at test time. This variant performs global pooling on the ResNet feature map after upsampling followed by a two layer MLP that predicts the viewpoint q and canonical pose. When training with this additional branch we find the camera-centered pose predictions show no improvement over baseline (Table 4). We also observe that the canonical pose predictions have higher error than the camera-centered predictions which is natural since the the model can't directly exploit the direct correspondence between the 2D keypoint locations and the 3D joint locations.

5.3 Comparison with state-of-the-art performance

Table 5 compares the proposed approach with the state-of-the-art performance on all 5 datasets. Note that our method is the first to evaluate 3d human pose estimation on the five representative datasets reporting both MPJPE and PCK3D, which fills in some blanks and serves as a useful baseline for future work. As can be seen, our method achieves state-of-the-art performance on H36M/GPA/SURREAL/3DPW/3DHP datasets in terms of MPJPE. While [10] uses additional data (both H36M and 3DHP, and LSP together with MPII) to train, they have slightly better performance on 3DHP in terms of PCK3D.

Qualitative Results: We visualize the prediction on the 5 datasets with model trained on H36M using our proposed method in Fig 7. The 2d joint prediction

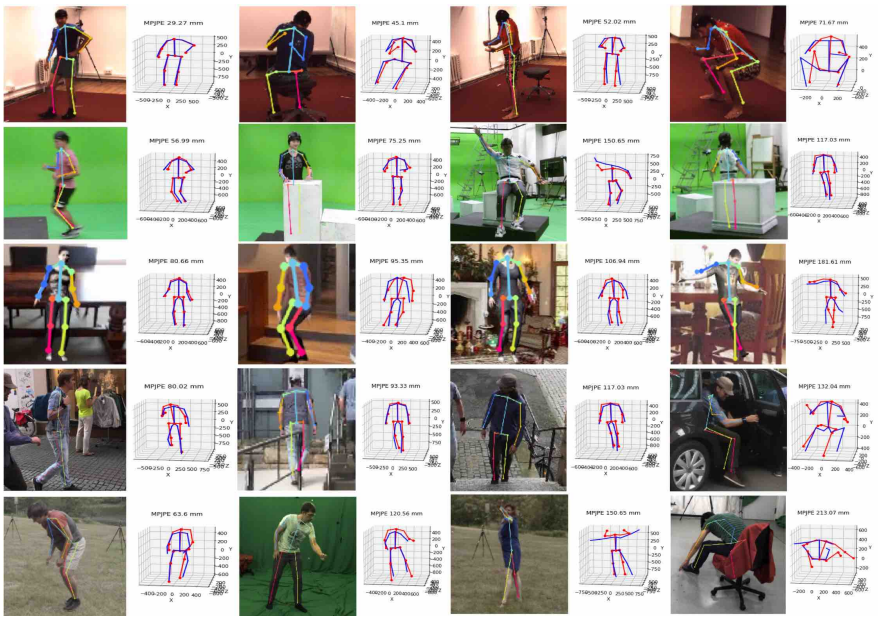


Fig. 7: Model predictiosn on 5 datasets from model trained on Human3.6M dataset. The 2d joints are overlaid with the original image, while the 3d prediction (red) is overlaid with 3d ground truth (blue). 3D prediction is **visualized in body-centered coordinate** rotated by the relative rotation between ground truth camera-centered coordinate and body-centered coordinate. From top to bottom are H36M, GPA, SURREAL, 3DPW and 3DHP datasets. We rank the images from left to right in order of increasing MPJPE.

is overlaid with cropped images while the 3d joint prediction is visualized in our proposed body-centered coordinates. From top to bottom are H36M, GPA, SURREAL, 3DPW and 3DHP datasets. We display the images from left to right in ascending order by MPJPE.

6 Conclusions

In this paper, we observe strong dataset-specific biases present in the distribution of cameras relative to the human body and propose the use of body-centered coordinate frames. Utilizing the relative rotation between body-centered coordinates and camera-centered coordinates as an additional supervisory signal, we significantly reduce the 3d joint prediction error and improve generalization in cross-dataset 3d human pose evaluation. Out model also achieves state-of-the-art performance on all same-dataset evaluations. We hope that our cross-dataset analysis is useful for future work and serves as a resource to guide future dataset collection.

7 Acknowledgement

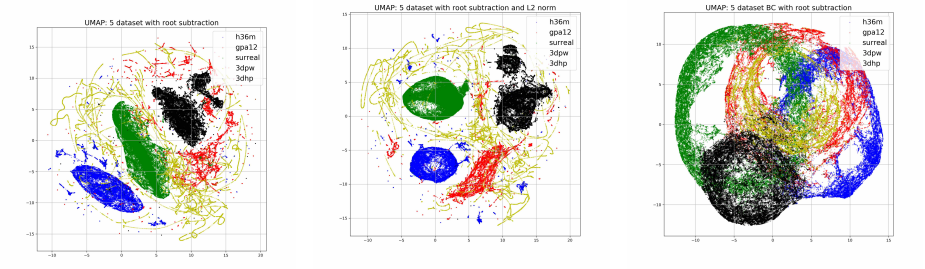
Acknowledgements: This work was supported in part by NSF grants IIS-1813785, IIS-1618806, and a hardware gift from NVIDIA.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
2. Arnab, A., Doersch, C., Zisserman, A.: Exploiting temporal context for 3d human pose estimation in the wild. In: CVPR (2019)
3. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
4. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2018)
5. khsanul Habibie, Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C.: In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In: CVPR (2019)
6. He, K., Gkioxari, G., Dollr, P., Girshick, R.: Mask r-cnn. In: CVPR (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
8. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. In: PAMI (2014)
9. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
10. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (2019)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
12. Kokkinos, I.: Ubernet: Training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: Arxiv (2016)
13. Lasinger, K., Ranftl, R., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In: Arxiv (2019)
14. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: aistats (2015)
15. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018)
16. Marinoiu, E., Zanfir, M., Olaru, V., Sminchisescu, C.: 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In: CVPR (2018)
17. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV (2017)
18. McInnes, L., Healy, J., Saul, N., Grossberger, L.: Umap: Uniform manifold approximation and projection. The Journal of Open Source Software (2018)

19. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3DV (2017)
20. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4460–4470 (2019)
21. Moon, G., Chang, J., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: ICCV (2019)
22. Nikos Kolotouros, Georgios Pavlakos, K.D.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019)
23. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: CVPR (2017)
24. Pavllo, D., Grangier, D., Auli, M.: Quaternet: A quaternion-based recurrent model for human motion. In: BMVC (2018)
25. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. In: TPAMI (2016)
26. Richter, S.R., Roth, S.: Matryoshka networks: Predicting 3d geometry via nested shape layers. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1936–1944 (2018)
27. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2d and 3d pose detection in natural images. In: PAMI (2019)
28. Shin, D., Fowlkes, C., Hoiem, D.: Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
29. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronizedvideo and motion capture dataset and baseline algorithm forevaluation of articulated human motion. In: IJCV (2010)
30. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV (2018)
31. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2088–2096 (2017)
32. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3405–3414 (2019)
33. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR (2011)
34. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: BMVC (2017)
35. Tulsiani, S., Gupta, S., Fouhey, D., Efros, A.A., Malik, J.: Factoring shape, pose, and layout from the 2d image of a 3d scene. In: CVPR (2018)
36. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Computer Vision and Pattern Regognition (CVPR) (2017)
37. Tung, H.Y.F., Tung, H.W., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture. In: CVPR (2009)
38. Umar Iqbal, Pavlo Molchanov, T.B.J.G.J.K.: Hand pose estimation via latent 2.5d heatmap regression. In: ECCV (2018)
39. Varol, G., Ceylan, D., Bryan Russell, a.J.Y., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3d human body shapes. In: CVPR (2019)

40. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR (2017)
41. Villegas, R., Yang, J., Ceylan, D., Lee, H.: Neural kinematic networks for unsupervised motion retargetting. In: CVPR (2018)
42. Wang, X., Cai, Z., Gao, D., Vasconcelos, N.: Towards universal object detection by domain attention. In: CVPR (2019)
43. Wang, Z., Chen, L., Rathore, S., Shin, D., Fowlkes, C.: Geometric pose affordance: 3d human pose with scene constraints. In: arxiv (2019)
44. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: ECCV (2018)
45. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV (2015)
46. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In: Advances in neural information processing systems. pp. 1696–1704 (2016)
47. Yin, X., Liu, X.: Multi-task convolutional neural network for pose-invariant face recognition. IEEE Transactions on Image Processing **27**(2), 964–975 (2017)
48. Yu, S., Yun, Y., Wu, L., Wenpeng, G., YiLi, F., Tao, M.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: ICCV (2019)
49. Zhang, X., Zhang, Z., Zhang, C., Tenenbaum, J., Freeman, B., Wu, J.: Learning to reconstruct shapes from unseen classes. In: Advances in Neural Information Processing Systems. pp. 2257–2268 (2018)
50. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: A weakly-supervised approach. In: ICCV (2017)
51. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images (2017)
52. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: ICCV (2019)



(a) UMAP with only root-subtraction (b) UMAP with root-subtraction and L2 normalization (c) UMAP body-centered coordinates with only root-subtraction

Fig. 8: Distribution of view-dependent, view-independent body-centered pose, visualized as a 2D embedding produced with UMAP [18].

Appendix

In the appendix, we **(1.)** visualize the UMAP embedding [18] of view-dependent pose (root-relate coordinates) from the five datasets. **(2.)** We provide results for other evaluation metrics (PMPJPE and PCK3D) that parallel the MPJPE results shown in the main paper. We also provide more detailed results showing the effectiveness of our quaternion loss in improving generalization of an alternate model of Zhou [50]. **(3.)** We visualize the distribution of viewpoints of five datasets in azimuth and elevation with cluster centers overlaid, **(4.)** We show selected examples based on the quaternion distribution pattern from five datasets. **(5.)** Finally, we show qualitative comparisons of training on each single dataset and testing across the five datasets, and training on five different datasets while testing on the same image from single dataset.

A UMAP Visualization

We visualize the UMAP [18] embedding of view-dependent coordinate (root-relate coordinate) of H36M [8], GPA [43] , SURREAL [40], 3DPW [15] and 3DHP [19] datasets in Fig 8a. We further normalize out skeleton size and visualize in Fig 8b. To compare with view-independent coordinate (body-center coordinate), we visualize them before L2 normalization in Fig 8c. We can see the body-centered, size normalized pose distribution (main paper) shows much higher overlap across datasets while the root-relative coordinates implicitly which encode camera orientation provide distinguishable information (dataset bias).

B PMPJPE, PCK3D results on posenet [21] and MPJPE results on Zhou [50]

We provide PMPJPE in Table 6 and PCK3d in Table 7 to demonstrate the effectiveness of adding quaternion loss to PoseNet [21]. To demonstrate the utility

		PA-MPJPE (in mm, lower is better)				
	Testing \Training	H36M	GPA	SURREAL	3DPW	3DHP
Baseline	H36M	**43.4**	75.0	69.6	91.3	75.0
	GPA	75.4	**41.7**	66.3	84.4	70.2
	SURREAL	76.5	73.5	**31.8**	85.8	77.9
	3DPW	68.0	66.9	64.3	**68.7**	68.1
	3DHP	88.5	91.2	86.9	111.3	**71.4**
Our Method	H36M	**42.5**	69.5	67.5	91.4	72.6
	GPA	71.4	**40.9**	65.6	81.4	70.6
	SURREAL	75.9	71.7	**31.7**	82.1	76.9
	3DPW	68.3	65.1	63.8	**65.2**	66.4
	3DHP	89.0	89.7	85.9	109.2	**70.6**
Same-Dataset Error Reduction ↓		0.9	0.8	0.1	3.2	0.8
Cross-data Error Reduction ↓		2.9	10.6	4.3	8.7	4.7

Table 6: Baseline cross-dataset test error and error reduction (Procrustese aligned MPJPE) from the addition of our proposed quaternion loss. Bold indicates the best performing model on each the test sets (rows). Blue color indicates test set which saw greatest error reduction.

		PCK3D (accuracy, higher is better)				
	Testing \Training	H36M	GPA	SURREAL	3DPW	3DHP
Baseline	H36M	**95.7**	75.7	52.3	70.6	77.8
	GPA	78.3	**96.3**	58.8	76.2	84.5
	SURREAL	76.4	84.5	**97.2**	73.6	81.0
	3DPW	83.2	78.7	54.5	**82.1**	81.7
	3DHP	76.1	70.3	44.8	68.4	**84.2**
Our Method	H36M	**96.0**	78.9	52.6	72.8	78.3
	GPA	81.5	**96.8**	59.3	76.4	84.8
	SURREAL	80.0	84.8	**97.3**	76.2	81.3
	3DPW	83.2	80.8	54.7	**84.6**	81.7
	3DHP	76.1	73.5	45.1	70.3	**84.3**
Same-Dataset Accuracy Increase ↑		0.3	0.5	0.1	2.5	0.1
Cross-data Accuracy Increase ↑		6.8	8.8	1.3	6.9	1.1

Table 7: Baseline cross-dataset test accuracy and accuracy increases (PCK3D) from the addition of our proposed quaternion loss. Bold indicates the best performing model on each the test set (rows). Blue color indicates test set which saw greatest accuracy increase.

of our quaternion loss on other models, we also show results based on retraining the model of [50] in Table 8 with MPJPE metric.

C Quaternion and cluster centers

Instead of colorizing each quaternion with cluster index, we directly visualize quaternion with the same color within each dataset in Fig 9, and also plot the cluster centers in the azimuth and elevation space.

D Sampled images from five datasets

Sampled images from H36M We sample images from the interesting azimuth/elevation pattern from H36M. We can see the images from Fig 10a are facing right while

		MPJPE (in mm, lower is better)				
	Testing \Training	H36M	GPA	SURREAL	3DPW	3DHP
Baseline	H36M	**72.5**	126.0	116.6	135.5	118.0
	GPA	110.5	**76.6**	97.3	116.2	100.6
	SURREAL	129.6	116.0	**54.1**	132.3	118.7
	3DPW	120.1	121.9	120.2	**108.5**	119.8
	3DHP	122.9	133.6	128.5	148.0	**104.5**
Our Method	H36M	**71.9**	122.2	115.4	134.4	109.9
	GPA	109.9	**72.4**	97.8	115.3	102.0
	SURREAL	129.2	113.5	**53.9**	126.5	119.4
	3DPW	119.1	119.3	119.9	**100.9**	116.5
	3DHP	122.5	130.2	127.6	145.7	**103.3**
Same-Dataset Error Reduction ↓		0.6	4.2	0.2	7.6	1.2
Cross-data Error Reduction ↓		2.4	12.3	1.9	10.1	9.3

Table 8: Retraining the model of Zhou *et al.* [50] using our viewpoint prediction loss also shows significant decrease in prediction error, demonstrating the generality of our finding.

images from Fig 10b are facing left. The index in the azimuth/elevation images corresponds with the index on top of images sampled and placed around the center figure.

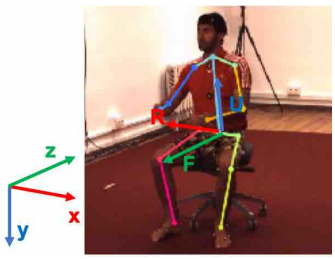
Sampled images from GPA/SURREAL We sample images from SURREAL and GPA with uniform azimuth from left to right, and place some randomness on elevation during sampling. We can see the patterns of sampled images from left to right: facing towards back and rotating to facing right, and facing towards the camera, and then facing back again in Fig 11.

Sampled images from 3DHP We sample images from 3DHP with uniform azimuth from left to right as shown in Fig 12b, uniform elevation from top to down as shown in Fig 12c, and from camera center as shown in Fig 12a, during sampling we add some randomness on sampled elevation/azimuth around camera centers.

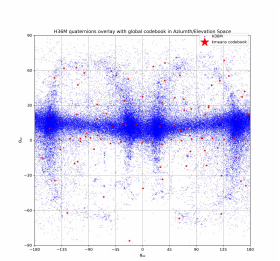
Sampled images from 3DPW We sample images from 3DPW with extreme elevation as shown in Fig 13a, and randomly as shown Fig 13b.

E Qualitative Results

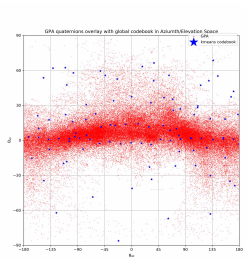
Qualitative Results trained on four datasets We visualize the prediction on the 5 datasets with model trained on **GPA**, **SURREAL**, **3DPW**, **3DHP** separately on using our proposed method in Fig 14,15,16,17. The 2d joint prediction is overlaid with cropped images while the 3d joint prediction is visualized in our proposed body-centered coordinates. From top to bottom are H36M, GPA, SURREAL, 3DPW and 3DHP datasets. We rank the images from left to right in MPJPE increasing order.



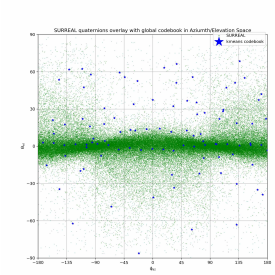
(a) Body-centered coordinate



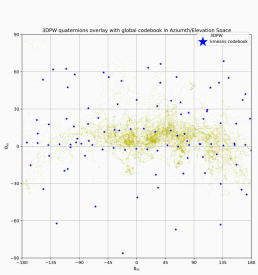
(b) H36M



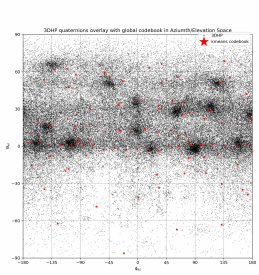
(c) GPA



(d) SURREAL



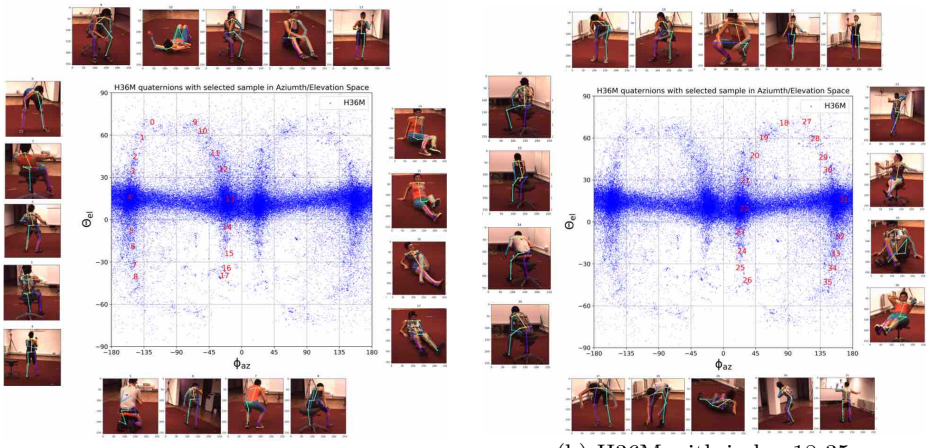
(e) 3DPW



(f) 3DHP

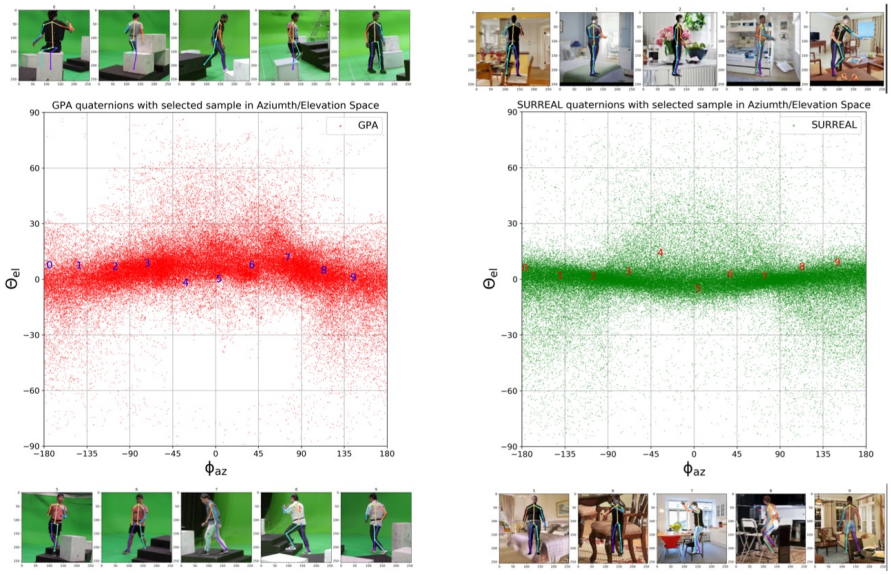
Fig. 9: **a**: Illustration of our body-centered coordinate frame (up vector, right vector and front vector) relative to a camera-centered coordinate frame. **b-f**: Camera viewpoint distribution of the 5 datasets overlaid with quaternion cluster centers. Quaternions (rotation between body-centered and camera frame) are sampled from training sets and clustered using k-means.

Qualitative Results tested on the same images We further visualize the models trained on 5 datasets, and test on images from the dataset H36M in Fig 18, GPA in Fig 19, SURREAL in Fig 20, 3DPW in Fig 21 and 3DHP in Fig 22. The results from left to right are models trained on H36M, GPA, SURREAL, 3DPW, and 3DHP. The RGB images are overlaid with 2d joint prediction from model trained on each dataset.



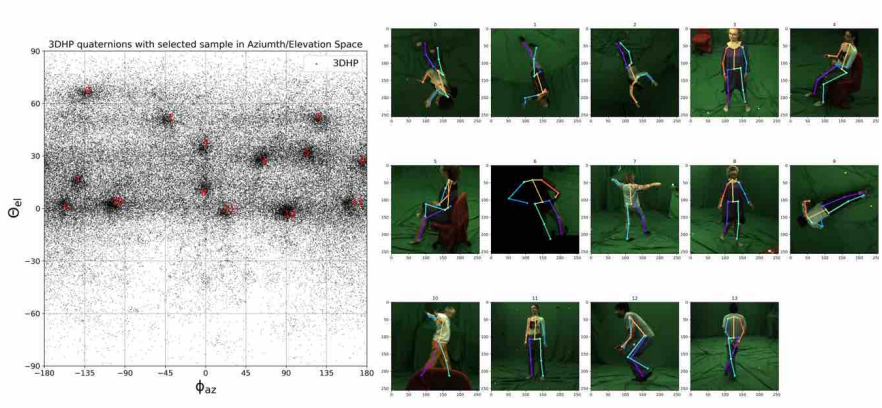
(a) H36M with index 0-17 (b) H36M with index 18-35

Fig. 10: H36M and sampled images.

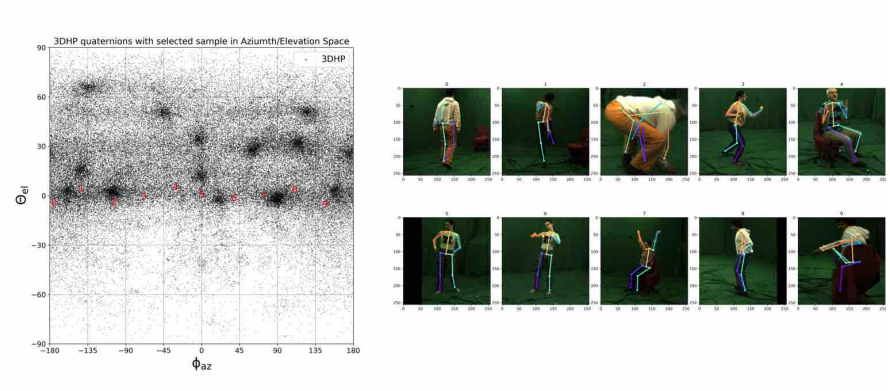


(a) GPA with sampled images. (b) SURREAL with sampled images

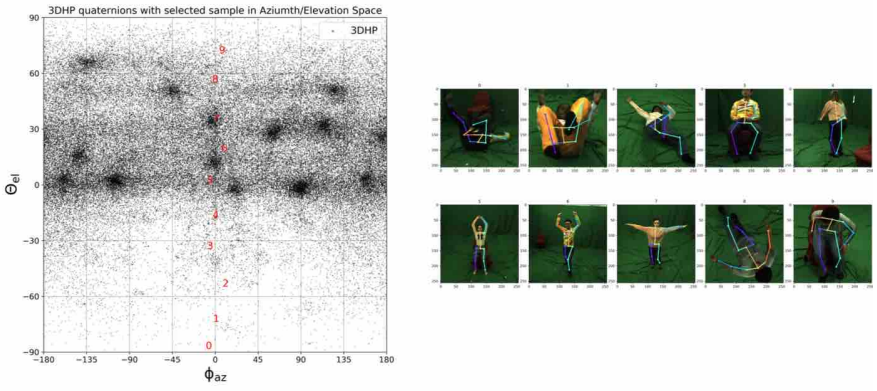
Fig. 11: GPA and SURREAL sampled images.



(a) 3DHP with images sampled from camera center.

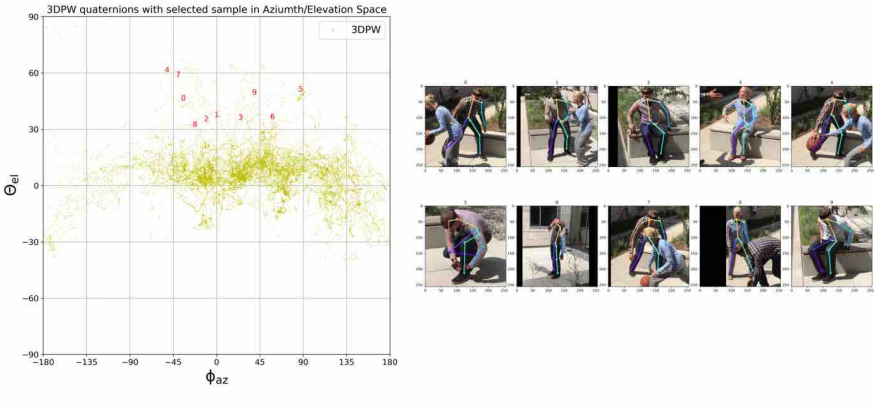


(b) 3DHP with sampled images in uniform azimuth space.

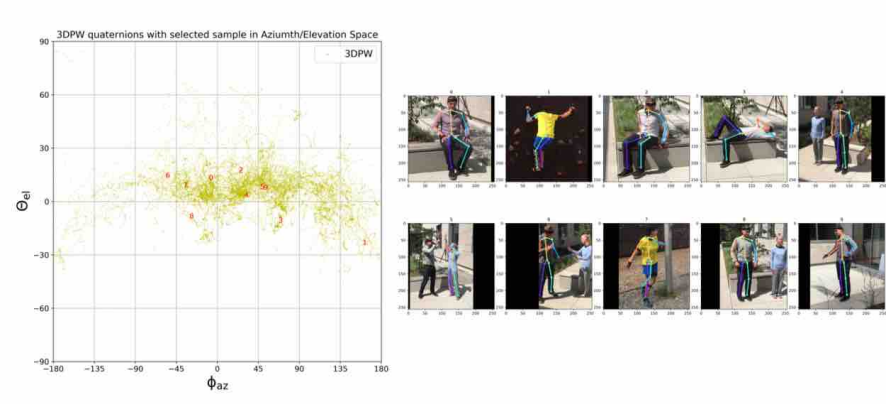


(c) 3DHP with sampled images in uniform elevation space.

Fig. 12: 3DHP sampled images.



(a) 3DPW with extreme elevation sampled images.



(b) 3DPW with random sampled images.

Fig. 13: 3DPW sampled images.

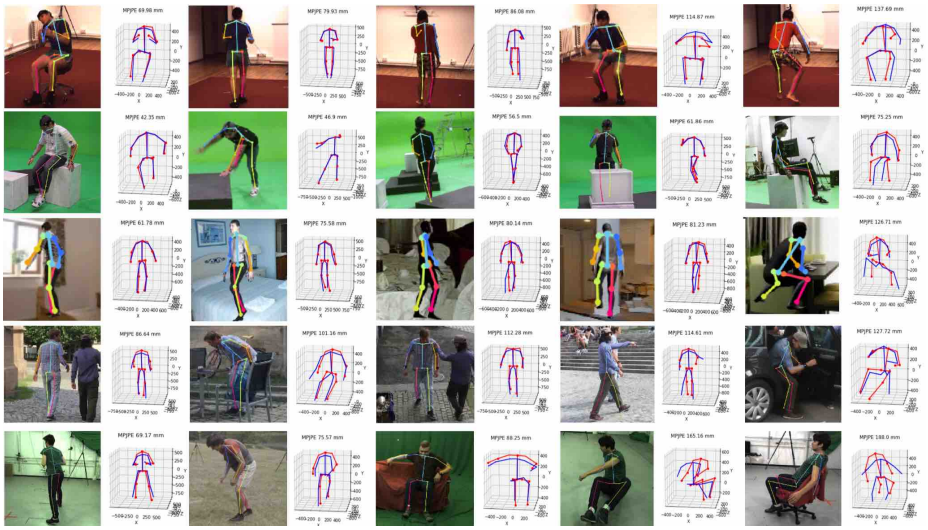


Fig. 14: Our prediction on 5 diverse dataset with model trained on GPA dataset. The 2d joints are overlaid with the original image, while the 3d prediction (red) is overlaid with 3d ground truth (blue). 3D prediction is **visualized in body-centered coordinate** rotated by the relative rotation between ground truth root-relative coordinate and body-centered coordinate. From top to bottom are H36M, GPA, SURREAL, 3DPW and 3DHP datasets. We rank the images from left to right in MPJPE increasing order.

Fig. 15: Our prediction on 5 diverse datasets with model trained on SURREAL dataset.

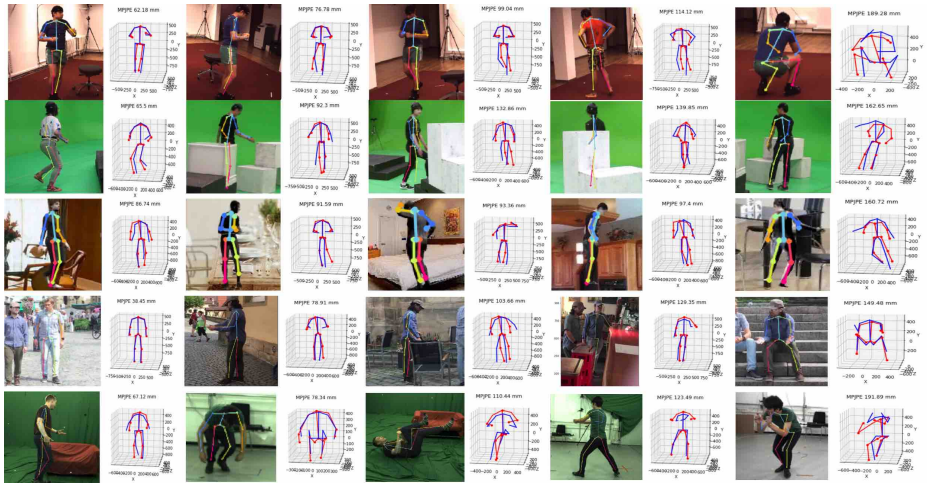


Fig. 16: Our prediction on 5 diverse datasets with model trained on 3DPW dataset.

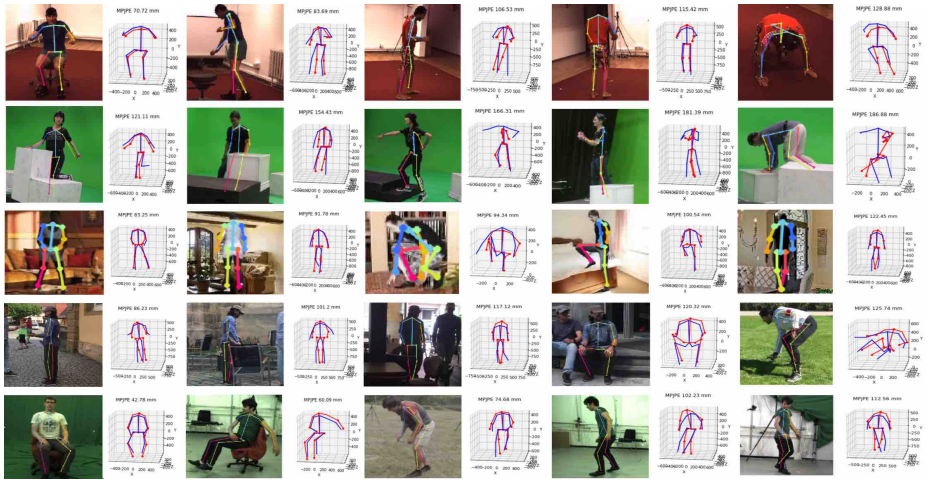


Fig. 17: Our prediction on 5 diverse datasets with model trained on 3DHP dataset.

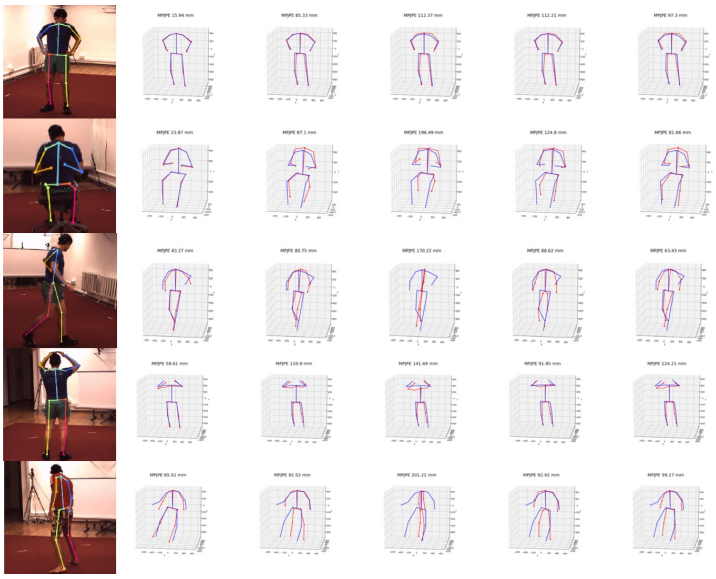


Fig. 18: Model trained on 5 models tested on the same images from H36M, from left to right (model trained on H36M, GPA, SURREAL, 3DPW, 3DHP).

Fig. 19: Model trained on 5 models tested on the same images from GPA, from left to right (model trained on H36M, GPA, SURREAL, 3DPW, 3DHP).

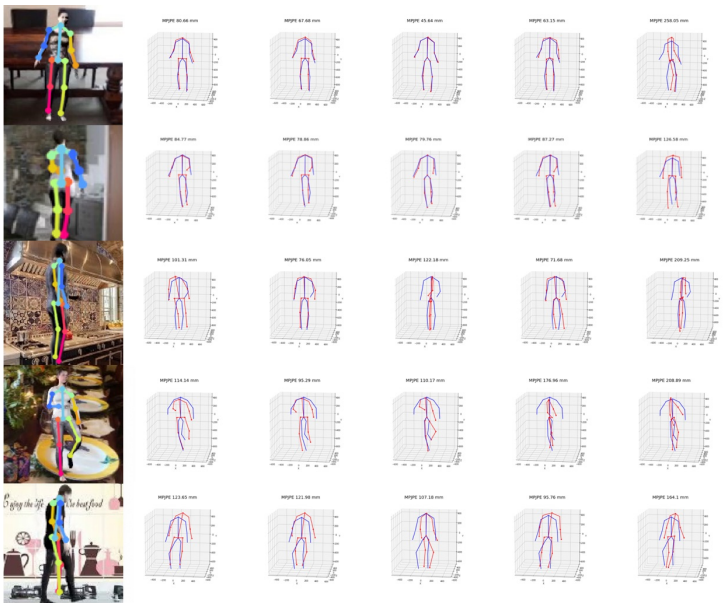


Fig. 20: Model trained on 5 models tested on the same images from SURREAL, from left to right (model trained on H36M, GPA, SURREAL, 3DPW, 3DHP).

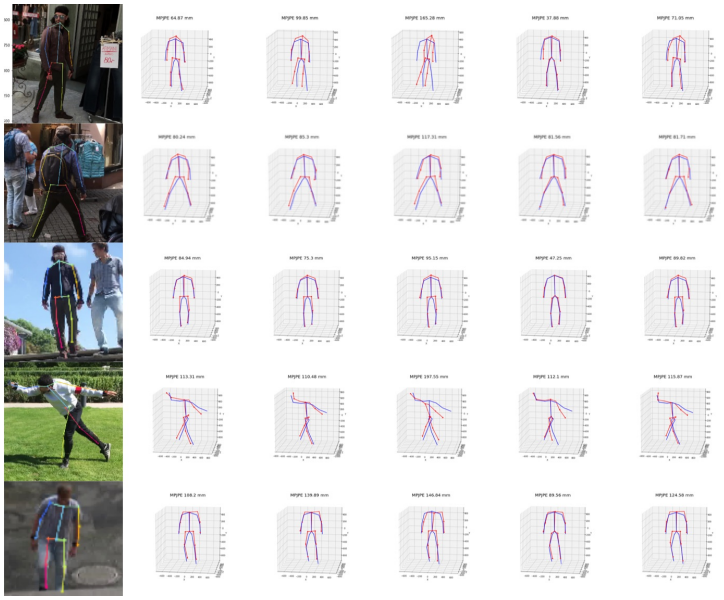


Fig. 21: Model trained on 5 models tested on the same images from 3DPW, from left to right (model trained on H36M, GPA, SURREAL, 3DPW, 3DHP).

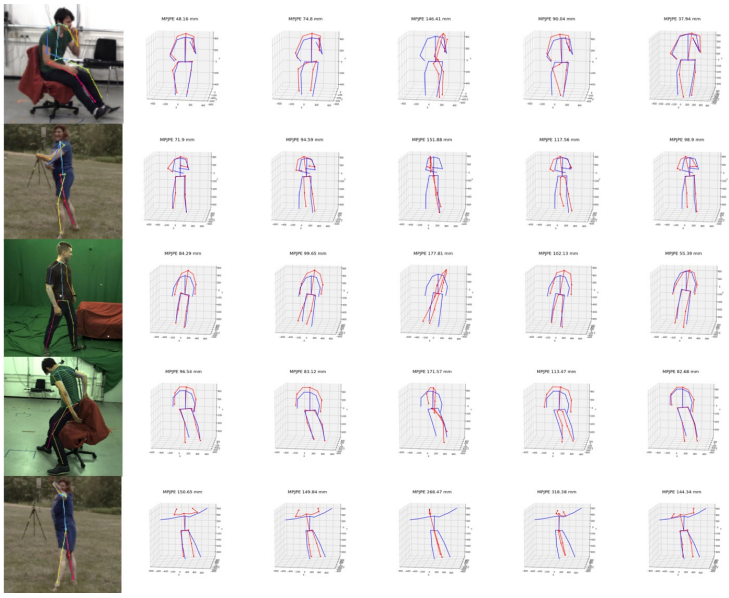


Fig. 22: Model trained on 5 models tested on the same images from 3DHP, from left to right (model trained on H36M, GPA, SURREAL, 3DPW, 3DHP).