Body Knowledge and Uncertainty Modeling for Monocular 3D Human Body Reconstruction
** Appendix **

Yufei Zhang[1], Hanjing Wang[1], Jeffrey O. Kephart[2], Qiang Ji[1]
[1]Rensselaer Polytechnic Institute, [2]IBM Research
{zhangy76, wangh36, jiq}@rpi.edu, kephart@us.ibm.com

In this Appendix, we first introduce additional details referred in the main manuscript which include

- Section A: the way of encoding the geometry constraints and its significance;
- Section B: additional introduction of the datasets and implementation;
- Section C: the method of quantifying the uncertainty of 3D mesh vertex prediction;
- Section D: implementation details of incorporating additional annotations.

Then we provide additional experiment results including

- Section E: example minority images in training and testing sets;
- Section F: additional qualitative evaluation;
- Section G: evaluation results of shape estimation;
- Section H: measuring the labeling noise presented in existing MoCap data through the proposed generic constraints.

A. Geometry Constraints

As illustrated in Figure 2b of the main manuscript, the shoulders, neck, and spine joints are coplanar; the hips and pelvis joints are collinear. Let $\mathbf{P}_i$ be the 3D position of joint i and $\mathbf{P}_{ij} = \mathbf{P}_j - \mathbf{P}_i$ be the bone vector, where $i, j = 0, ..., 6$ and $i \neq j$. The geometry constraints can be imposed via encouraging the angle between bone $\mathbf{P}_{64}$ and $\mathbf{P}_{65}$ to be 180 degrees, and the angle between bone $\mathbf{P}_{02}$ and the norm of plane $\mathbf{P}_{0,1,3}$ to be 90 degrees. The losses can be formulated accordingly as

$$\mathcal{L}_{coplanar} = \frac{|(\mathbf{P}_{01} \times \mathbf{P}_{03}) \cdot \mathbf{P}_{02}|}{\|\mathbf{P}_{01} \times \mathbf{P}_{03}\| \|\mathbf{P}_{02}\|}, \tag{1}$$

$$\mathcal{L}_{collinear} = \frac{|\mathbf{P}_{64} \times \mathbf{P}_{65}|}{\|\mathbf{P}_{64}\| \|\mathbf{P}_{65}\|}, \tag{2}$$

$$\mathcal{L}_{geometry} = \mathcal{L}_{coplanar} + \lambda_{colinear} \mathcal{L}_{collinear}. \tag{3}$$

The geometry constraints are derived based on the knowledge of human body structure. In Table 1 row 2, we demonstrate that imposing the geometry constraints ensures realistic upper body reconstruction (the reconstructed joints becomes colinear and coplanar). Moreover, these geometry characteristics can be exploited to solve the inherent depth ambiguity in lifting 2D observation to its 3D configuration. Specifically, let $\mathbf{P}_1$, $\mathbf{P}_2$, and $\mathbf{P}_3$ be the 3D position of three collinear points in the camera coordinate system. Under perspective projection, we have

$$\lambda \mathbf{p}_i = \mathbf{K} \mathbf{P}_i, \tag{4}$$

where $i = 1, 2, 3$, $\mathbf{K}$ is the camera intrinsic matrix, and λ is a scalar. For Equation 4, $\mathbf{P}_i$ can not be uniquely solved due to the depth ambiguity (the 2D-3D correspondences provide two equations but with three unknowns). However, when the three points are collinear, two additional equations can be introduced through

$$\frac{\mathbf{P}_1 - \mathbf{P}_2}{\|\mathbf{P}_1 - \mathbf{P}_2\|} = \frac{\mathbf{P}_1 - \mathbf{P}_3}{\|\mathbf{P}_1 - \mathbf{P}_3\|}. \tag{5}$$

Given the camera intrinsic parameters and the 3D distance between the three points, unique values of $\mathbf{P}_1$, $\mathbf{P}_2$, and $\mathbf{P}_3$ can be solved. Similarly, the coplanarity can introduce additional constraints for alleviating the depth ambiguity. As we do not assume the camera information is given, further study of the geometry constraints is not discussed here.

B. Datasets and Implementation Details

Datasets. H36M includes 5 subjects performing 15 daily actions like Eating, Greeting, and *et al*, and it consists of a

Models	Constraint Satisfaction					Reconstruction Error	
	Anatomy		Biomechanics		Physics	MPE	P-MPE
	Bone	Geometry	Angle	Angle-inter			
$\mathcal{L}_{NLL}$	101.2	83.2/166.5	30.7/30.8	109.1	97	296.5	161.2
$\mathcal{L}_{NLL} + \mathcal{L}_{geometry}$	130.8	89.4/178.7	41.6/41.7	35.2	100	415.3	311.5
3DPW GT	21.0	89.8/178.6	4.1/4.1	6.6	2	-	-

Table 1. **Evaluation of the constraints satisfaction on the ground truth data and the usage of the geometry constraints.** For quantifying the constraint satisfaction, we compute the mean per bone length error (Bone, in mm), average angle induced by the coplanar/colinear joints (Geometry, in degrees), mean per joint angle violations with/without considering the inter-joint dependency (Angle/Angle-inter, in degrees), and percentage of data with penetration (Physics, in percents). The bone and geometry constraints are soft constraints, while the biomechanic and physic constraints are hard constraints that should be strictly satisfied. The angle induced by the coplanar and coplinear joints should approximate to 90 and 180 degrees, respectively.

total of 312,188 training images. MPI-3D includes 8 subjects covering 8 typical action classes like Exercise, Sitting, and *et al.*, with a total of 96,620 valid training images. COCO is a dataset widely used for segmentation and detection tasks. LSP and LSP-Extended (10,482 images) and MPII (14,806 images) are standard datasets for 2D pose estimation and involve more diverse poses than COCO.

Implementation. Our implementation uses Pytorch. The model is trained using Pytorch's Adam solver with a learning rate of 10^{-5} and weight decay 10^{-4}. The training is conducted on one 2080Ti GPU with batch size of 64. The images from different datasets are fed into one mini-batch with the following split: H36M (0.35), MPI-INF-3DHP (0.1), COCO (0.35), MPII (0.1), and LSP and LSP-Extended (0.1). During training, we observe that it is efficient to train the regression model by first encoding the generic prior. We hence first train the regression model on H36M [1] for $\sim 150K$ iterations and then continue training on all the datasets for $\sim 500K$ iterations. Once the initial model is trained, we quantify the uncertainty of all the training samples and compute the corresponding uncertainty-guided refinement weights. Based on the computed weights, we further refine the initial model for $\sim 100K$ iterations and obtain the final model.

Regarding the hyperparameters, we use 10 for the keypoints reprojection loss, 500 for the body anatomy loss, 1000 for the the biomechanics loss, 1000 for physic loss, and 1 for scaling the refinement weights. We also add regularization on the trace of the predicted covariance matrix of the 2D keypoint projection with a weight of 50. This term encourages the model to converge to the position with small 2D projection error.

When calculating the model memory, we use Pytorch's `model.parameters()` and `model.buffers()` to count all the parameters and buffers stored in a model. When evaluating the model speed, we perform the model inference on a computer with a Intel(R) Xeon(R) W-2135 CPU and one 2080Ti GPU.

C. Uncertainty Quantification for 3D Mesh Vertex Prediction

Not limited to quantifying the uncertainty of 2D body keypoint prediction, KNOWN can also quantify the epistemic uncertainty of the 3D vertex prediction. Specifically, for 3D body mesh vertices $\mathbf{M}$, its conditional probability given 3D body model parameters $\mathbf{Y}$ follows Gaussian distribution:

$$p(\mathbf{M}|\mathbf{Y}) = \mathcal{N}\big(\boldsymbol{\mu}_{\mathbf{M}}(\mathbf{Y}), \boldsymbol{\Sigma}_{\mathbf{M}}(\mathbf{Y})\big), \tag{6}$$

where the mean of the Gaussian distributions are specified by the body pose and shape parameters via the forward kinematic process. We quantify the epistemic uncertainty of the 3D vertex prediction as

$$\underbrace{\mathrm{Cov}_{p(\mathbf{Y}|\mathbf{X};\mathbf{W})}\left[\mathrm{E}_{p(\mathbf{M}|\mathbf{Y})}[\mathbf{M}]\right]}_{\text{Epistemic uncertainty}} = \mathrm{Cov}_{p(\mathbf{Y}|\mathbf{X};\mathbf{W})}\left[\boldsymbol{\mu}_{\mathbf{M}}(\mathbf{Y})\right]. \tag{7}$$

Directly computing the right side of Equation 7 is difficult. We approximate the value via sample covariance:

$$\underbrace{\mathrm{Cov}_{p(\mathbf{Y}|\mathbf{X};\mathbf{W})}\left[\mathrm{E}_{p(\mathbf{M}|\mathbf{Y})}[\mathbf{M}]\right]}_{\text{Epistemic uncertainty}} \approx \mathrm{Cov}\left[\{\boldsymbol{\mu}^s_{\mathbf{M}}\}_{s=1}^S\right], \tag{8}$$

where $\{\boldsymbol{\mu}^s_{\mathbf{M}}\}_{s=1}^S$ is computed using the samples of $\mathbf{Y}$.

For the visualization of the 3D vertex prediction uncertainty in Figure 5 of the main manuscript, the vertex color represents the epistemic uncertainty computed following the method introduced above. The colors are obtained from a standard color map after normalizing the scalar uncertainty value of each vertex to a range from 0 to 1.

Compared to the uncertainty quantified on the 2D keypoint projections, the epistemic uncertainty quantified on 3D vertex prediction does not involve the projection process. Future work can consider further distinguish these two types of uncertainty to account for the uncertainty in estimating the camera parameters.

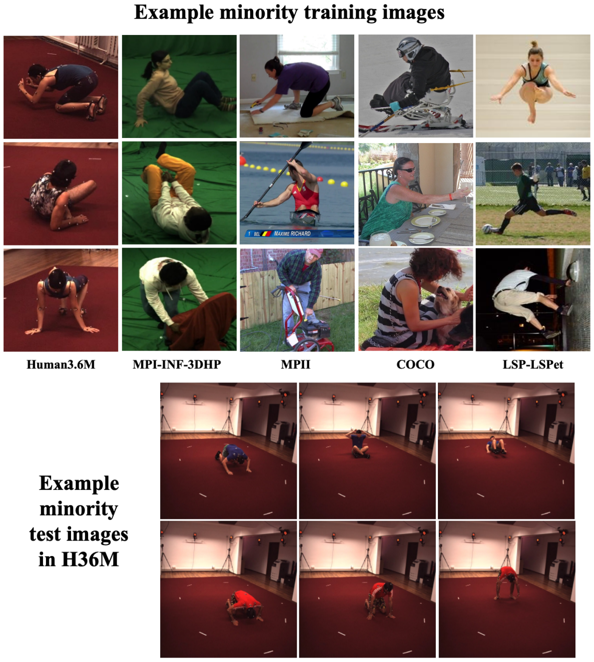

Example minority training images
Human3.6M
MPI-INF-3DHP
MPII
COCO
LSP-LSPet
Example minority test images in H36M

Figure 1. **Example minority images.**

D. Utilizing Additional Annotations

Training of KNOWN can easily incorporate annotations from different sources when they are available. Specifically, additional annotations can be incorporated during model finetuning via minimizing a corresponding reconstruction error. We here discuss the usage of the paired 3D annotations.

Paired 3D annotations indicate either 3D body joint position annotation or 3D body model parameter annotation that are paired with an image. Paired 3D annotations are hard to obtain and they are typically collected indoors. For the employed training datasets, H36M, MPI-3D, COCO, MPII, and LSP and LSP-Extended, only H36M and MPI-3D include the 3D body joint position annotations. To incorporate these annotations, we formulate the following loss functions:

$$\mathcal{L}_{pair3D} = \|\mathbf{P} - \hat{\mathbf{P}}(\boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\beta)\|_2^2, \tag{9}$$

where $\hat{\mathbf{P}}(\boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\beta)$ is the predicted 3D body joint position computed based on the mean pose and shape estimates. The overall training loss is then calculated as:

$$\mathcal{L} = \sum_{i=1}^{N} w_i(\mathcal{L}_{NLL,i} + \mathcal{L}_{pair3D}) + \mathcal{L}_{generic,i}. \tag{10}$$

The overall training loss includes the uncertainty-guided refinement weights to effectively leverage the data from a specific domain based on their uncertainty.

LSP	Parts	
	Acc.	F1
HMR [2]	87.00	0.59
Ours	**87.40**	**0.65**

Table 2. **Evaluation of body shape estimation.** The accuracy (Acc.) and F1 score (F1) are computed on six-part body segmentation on LSP's test set.

E. Example Minority Images

In Figure 1, we present example minority images from different training datasets and the testing set of H36M (Protocol 2). The minorities are mainly the images with large camera angle, severe occlusion, or extreme poses — situations that make their reconstruction particularly challenging. KNOWN successfuly improve model performance on these challenging image via uncertainty-guided refinement.

F. Additional Qualitative Evaluation

In Figure 2, we present additional qualitative evaluation on the majorities (small epistemic uncertainty) and minorities (large epistemic uncertainty). The majorities in each test set are the images with large data density. As shown, the majorities typically posses simple poses and few occlusion. The model performance with and without employing the refinement are similar on the majorities. By contrast, the minorities in each test set are the images with low data density and they always contain more challenging poses and severe occlusion. Applying the uncertainty-guided refinement loss shows significant improvements on the minorities. Specifically, the figures at the last row of Figure 2 are the evaluation on a minority image from LSP's test set. The left leg and the two arms of the person in the input image are occluded or blurred with the backgrounds. As a result, our model shows large epistemic uncertainty on these regions. Moreover, utilizing the uncertainty-guided refinement improves the model performance on these challenging cases, such as the left arm aligns better with the image.

G. Additional Quantitative Evaluation on Body Shape Estimation

We demonstrate KNOWN's improved body shape estimation performance via six-part body segmentation accuracy evaluated on the LSP test set, following the typical evaluation protocol used by [2]. During evaluation, body part segmetations are obtained by rendering the 3D prediction on image using the predicted camera parameters. Without using any 3D information, KNOWN's body part segmentation retains accuracy of 87.40 and F1 score of 0.65, which are better than HMR's 87.40 and 0.59, respectively.

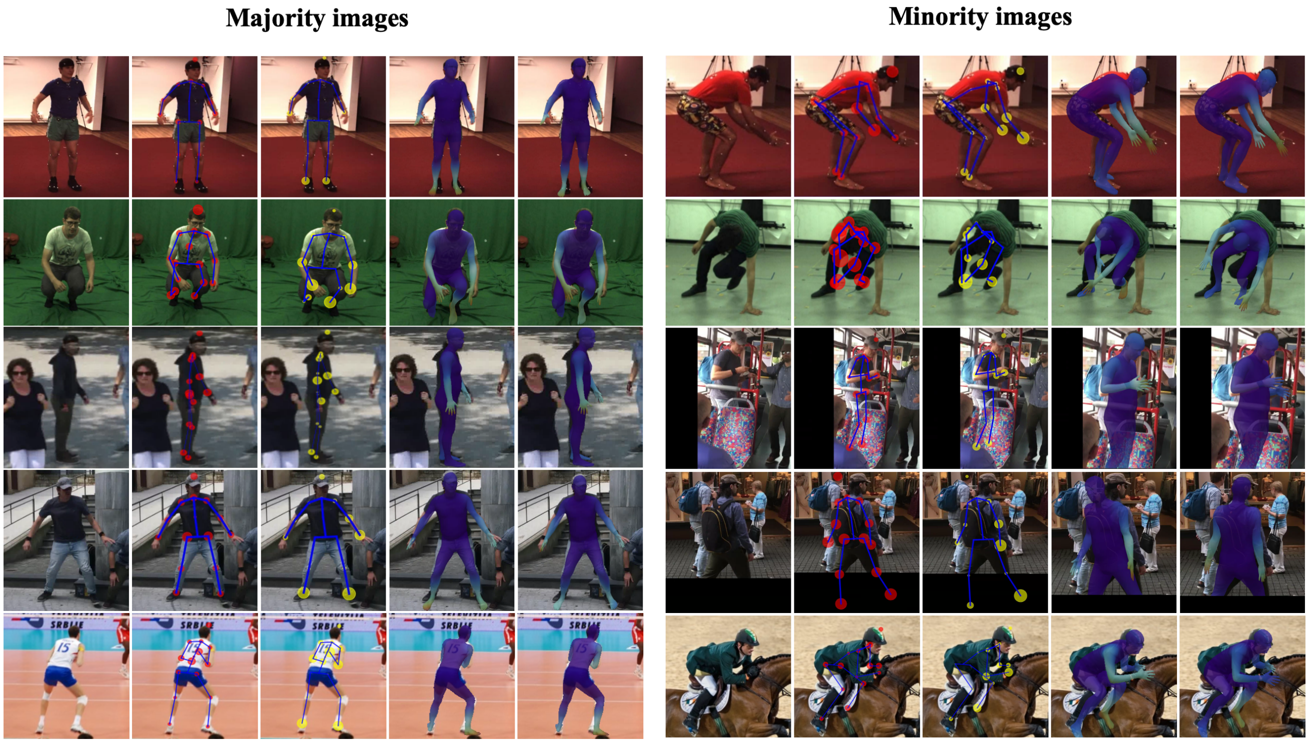
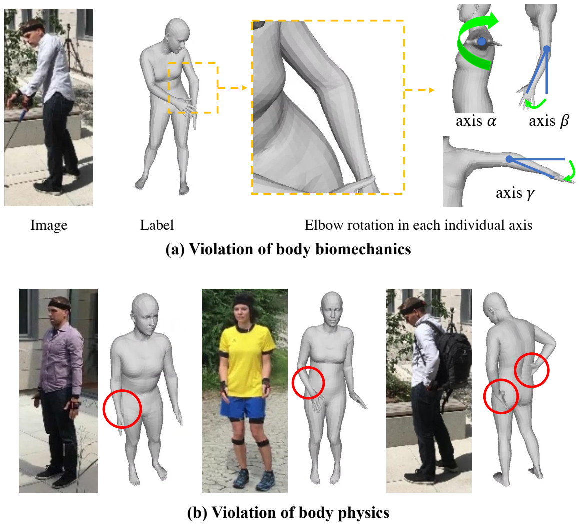


Figure 2. **Qualitative evaluation on the majorities and minorities.** The test images from top to bottom are from H36M (row 1), MPI-3D (row 2), 3DPW (row 3-4), and LSP (row 5), respectively. The images from lift to right are the input image, data uncertainty (without refinement), model uncertainty (without refinement), and 3D reconstruction results (without/with refinement), respectively.

H. Measuring Labeling Noise Presented in Existing MoCap Datasets

The proposed generic constraints can be used to measure the data noise due to violation of the physical constraints. Example noisy data occurred in existing datasets is shown in Figure 3. In details, in Figure 3 (a), the SMPL model annotation is consistent with the original image in general, while the pose label of the left elbow shows infeasible bending. Specifically, for this data sample, α, β, and γ at left elbow is -6.4, -24.2, and -18.2 degrees, respectively. While the corresponding valid angle ranges are $(-180, 90)$, $(-166, 0)$, and $(0, 0)$. The rotation defined by γ clearly violates the biomechanic constraint and leads to an unrealistic configuration at the left elbow. Furthermore, example violations of body physics are visualized in Figure 3(b). Similarly, the overall configuration given by the label is consistent with the original image but has penetration between body parts, including the unrealistic contact between body hand and torso. These violations of body biomechanics and physics mainly stem from the label generation process, where the labels are generated by fitting to only a set of sparse 3D markers [3, 4]. Although the annotations are generally aligned with the corresponding image data, the violation of the hard generic body constraints



(a) Violation of body biomechanics

(b) Violation of body physics

Figure 3. **Labeling noise presented in existing MoCap data.** The labels in existing MoCap datasets are generated via fitting a parametric body model to a set of sparse 3D markers, which can lead to the violation of (a) body biomechanics; and (b) body physics (the body parts with penetration are marked with red circles) constraints.

leads to unrealistic 3D configuration. We propose to impose the generic body constraints to ensure more physically plausible 3D reconstruction and avoid being affected by the labeling noise.

References

[1] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2

[2] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 3

[3] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014. 4

[4] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 4