Supplementary Material: Dynamic Multi-Person Mesh Recovery From Uncalibrated Multi-View Cameras

Buzhen Huang Yuan Shu Tianshu Zhang Yangang Wang
Southeast University, China

1. Datasets

Campus and Shelf [1] The Campus and Shelf datasets contain more than 3 characters with partial occlusions and across view ambiguities. We follow the same evaluation protocol as in previous works [3] and compute the PCP (percentage of correctly estimated parts) scores to measure the accuracy of 3D pose estimation.

Panoptic [6] This dataset is captured in a studio with 480 VGA cameras and 31 HD cameras, which contains multiple people engaging in social activities. We conduct qualitative and quantitative experiments on *160906_pizza1* sequence with HD cameras.

MHHI [9] is a multi-person dataset that contains complex and extreme poses as well as fast motion. The Fight sequence is publicly available and captured in a marker-based manner. For a fair comparison, the quantitative experiments are conducted on this sequence.

AMASS [10] is a large collection of 15 motion capture datasets with a unified SMPL representation. The dataset is used for motion prior training.

3DOH [14] is a 3D human occlusion dataset. We qualitatively evaluate our method on this dataset to reveal the superiority of our approach on motion capture from occluded scenarios.

Human3.6M [5] is a large-scale, single human dataset captured in a controlled scene, which consists of 11 subjects with 4 views. It provides accurate 3D joint positions and camera parameters. We follow [4, 7] to use S9 and S11 for evaluation.

MPI-INF-3DHP [11] is captured using a multi-view system. The standard testset contains 1 view under indoor and outdoor scenes. We use the testset to evaluate our method in the single-view setting.

2. Generative Model Evaluation

We conducted an experiment to demonstrate the generative ability of our motion prior. Randomly sample the latent code for each frame will lead to an incoherent motion. Thus, we sampled the latent code of the start and end frames from the standard Gaussian distribution and generated the code of the entire motion with linear interpolation. Fig.1 shows the comparison of the motion prior between with and without local linear constraint. Without the constraint, the interpolated latent code will produce preternatural interpenetration. On the contrary, with the local linear constraint, we can generate temporal coherent and diverse motions by linearly interpolating the sampled latent code.

Method	*Human3.6M*			*MPI-INF-3DHP*		
	PA-MPJPE	MPJPE	Accel	PA-MPJPE	MPJPE	Accel
*Li *et al.* [7]	43.8	64.8	–	65.1	97.6	–
*Liang *et al.* [8]	45.1	79.9	–	–	62.0	–
Huang *et al.* [4]	47.1	58.2	–	–	–	–
VPoser-t	34.7	53.5	10.2	73.5	103.6	105.4
w/o local linear	32.5	44.2	7.1	64.7	99.1	43.7
Ours	30.3	43.3	4.6	62.4	90.2	32.3

Table 1: Comparison on single-person datasets. The MPJPE and PA-MPJPE are in mm. The Accel is acceleration error in mm/s^2. Our method gets more accurate and temporal coherent results. * denotes learning-based regression.

3. More Results on Single-person Datasets

3.1. Results

[7] and [8] train a neural network to regress SMPL parameters from multi-view images. Huang *et al.* [4] proposed an optimization-based method to fit the human model to multi-view 2D keypoints. We compared our method with these baselines on single-person datasets. As shown in Tab.1, we reported the mean per joint position error (MPJPE), the MPJPE after rigid alignment of the prediction with ground truth using Procrustes Analysis (PA-MPJPE) to evaluate the accuracy of estimated skeleton joints. Furthermore, we used the acceleration error (Accel), which is calculated as the difference in acceleration between the ground-truth and predicted 3D joints, to describe the quality of the predicted motion. The results in Tab.1 demonstrate that our method achieves state-of-the-art. Besides, with the local linear constraint, the acceleration error decreases by 2.5 on Human3.6M dataset, proving that it produces more coherent motions. We then conducted a qualitative comparison between VPoser-t and our method in Fig.2. The first row of Fig.2 shows a single view sam-

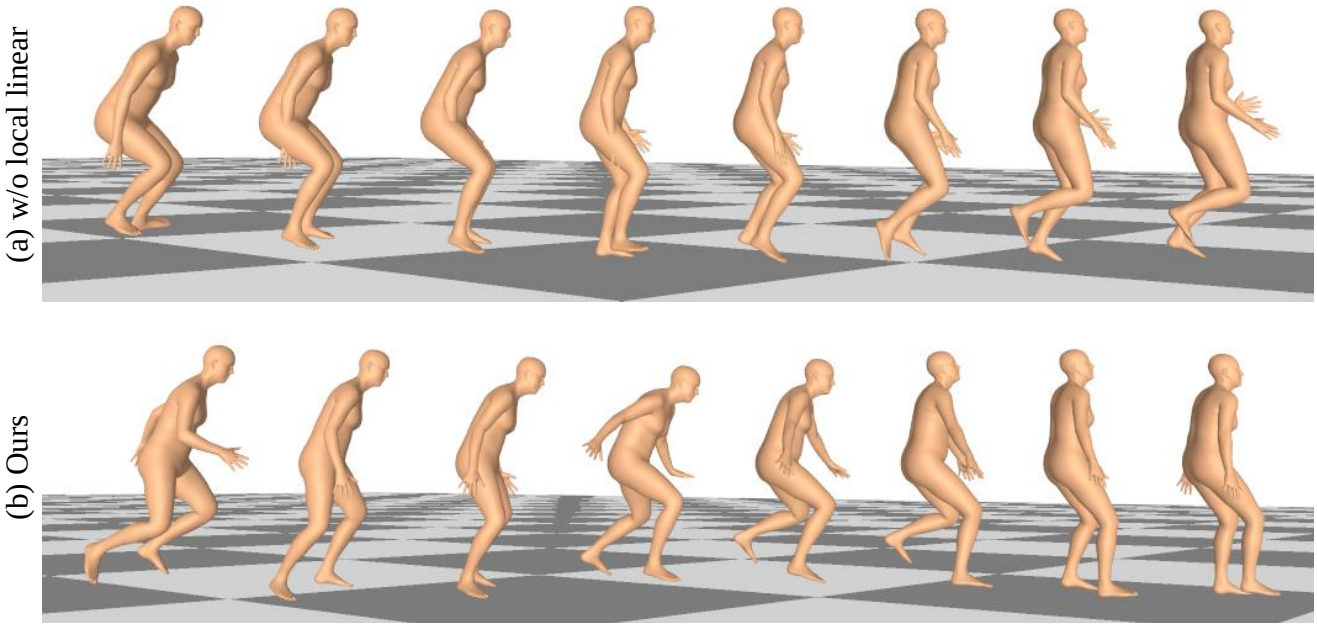
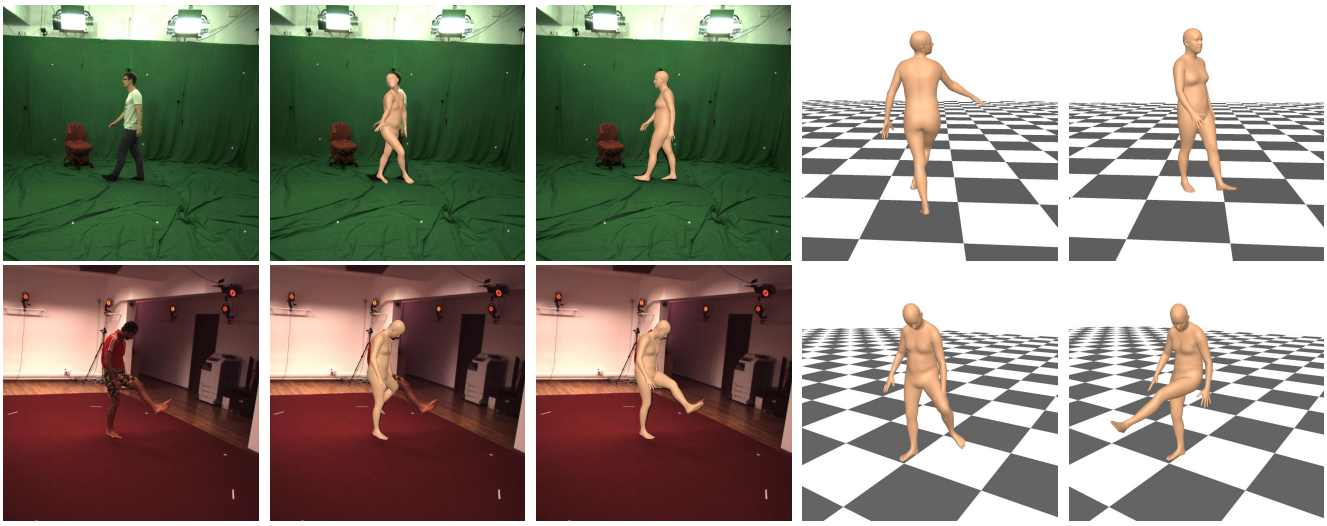


Figure 1: Without the local linear constraint, the motion prior produces preternatural interpenetration on the meshes.



(a) RGB image (b) Overlay VPoser-t (c) Overlay ours (d) Results VPoser-t (e) Results ours

Figure 2: Row 1 shows a single-view case of a side view person, and row 2 is a multi-view case with the inaccurate detected 2D poses. Our method is more robust to the noises and ambiguities than VPoser-t.

ple on MPI-INF-3DHP dataset. Due to the ambiguity of symmetrical skeleton joints on the side view, VPoser-t can not use temporal information to penalize incoherent results. However, our motion prior encodes the global dynamics and local kinematics and can produce a natural mesh. In the second row, the inaccurate 2D poses result in a jitter of 3D mesh for VPoser-t. The results demonstrate that our method is more robust to the noises.

4. Details

4.1. Data augmentation

Due to the limited human motion data, we use data augmentation to enhance the generalization performance of the model when training the motion prior. The strategy mainly includes 1) Upsampling and downsampling. We upsample or downsample the origin sequences to generate motions in different frame rates. 2) Reverse sampling. We sample the sequence from the end frame to the start frame to generate a new sequence. 3) Flip sampling. Since the human

body is symmetrical, we generate new motions by following the kinematic tree of the human model to mirror the motion across the left and right.

4.2. Camera Initialization

The initial extrinsic parameters of cameras are of great importance to both denoising and joint optimization. To ensure that the coarse values are in a reasonable range, a re-projection error is used to judge the cameras. We first calculate the 3D skeleton joints from different view 2D poses with the initial camera parameters. We then project the 3D joints to different views. If the intersection of union (IoU) of the bounding-boxes of the projected 2D joints and detected 2D poses is above 0.4, we consider the initial estimated camera parameters are reliable. Otherwise, we estimate other extrinsic parameters in the next frame. We use the reliable results as the initial value for denoising and joint optimization.

4.3. Physics-geometry Consistent Denoising

A semi-positive definite matrix $\mathcal{M}$ can ensure that the correspondences between different views are correct, thus reducing the influence of the noises. During the iteration, the $\mathcal{M}$ is a real matrix whose element values are in the range of 0 - 1. The alternating direction method of multipliers (ADMM) [2] is adopted to solve this problem. We select the result with the most corresponding views from $\mathcal{M}$ as the final output and use the selected result for joint optimization. The remaining results without correspondences are the wrong detections.

4.4. Joint Optimization

Initialization. Although the motion prior is compact, jointly optimizing large-scale multi-person motion sequences and camera parameters is still a highly non-convex problem. To reduce the solution space, we use the initial camera parameters estimated in Sec.4.2 to initialize the global positions and rotations of human models in each frame. The coarse 3D skeleton joint positions are first triangulated. Then we rigidly align the models to the estimated 3D joints. The rotations and translations of the aligned models are used as initial values for the joint optimization.
Optimization. Our optimization is implemented in PyTorch [13] using L-BFGS [12] optimizer. On a desktop with an Intel(R) Core(TM) i9-9900K CPU and a GPU of NVIDIA GeForce RTX 2080Ti, 30s 5-view RGB videos with 4 people take about 8.5 min to fit, in which the physics-geometry consistent denoising takes about 8s and joint optimization spends 8.4min.

5. Limitations and Future Works

This work critically relies on the initial extrinsic camera parameters. Although Sec.4.2 uses re-projection to filter out unreliable results, it is still hard to be applied when the view does not contain the target person. The learning-based camera prediction may be used to estimate more robust initial values for future works. It is also an interesting direction to utilize human cues to estimate intrinsic parameters, and provide a complete pipeline for calibration and distortion correction. Furthermore, due to the coupling of human dynamics and camera motions, the proposed method can only reconstruct human meshes with fixed cameras. In the future, we will add a physical prior to the cameras and build independent view-view and model-view correspondences to decouple camera motions and human dynamics, thus performing calibration and mesh recovery from large-scale scenarios.

References

[1] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, 2014. 1

[2] S. Boyd, N. Parikh, and E. Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011. 3

[3] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *CVPR*, 2019. 1

[4] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 international conference on 3D vision (3DV)*, pages 421–430. IEEE, 2017. 1

[5] C. Ionescu, D. Papava, V. Olaru, and C. 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. 1

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

[7] Z. Li, M. Oskarsson, and A. Heyden. 3d human pose and shape estimation through collaborative learning and multi-view model-fitting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1888–1897, 2021. 1

[8] J. Liang and M. C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1

[9] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR*, 2011. 1

[10] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 1

[11] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose esti-

mation in the wild using improved cnn supervision. In *3DV*, 2017. 1

[12] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006. 3

[13] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 3

[14] T. Zhang, B. Huang, and Y. Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 1