HardMo: A Large-Scale Hardcase Dataset for Motion Capture

Supplementary Material

Overview

In this supplementary material, we present more details and more experimental results that are not included in the main paper. The contents include:

- More experiment setup details in Sec. S-A.
- The metrics used for the evaluation in Sec. S-B.
- Evaluation on Automatic Annotation Pipeline in Sec. S-C.
- Evaluation on 3DPW and LSPet in Sec. S-D.
- Impact of Data-Scale in Sec. S-E.
- More details about Collection Process in Sec. S-F.
- Additional Qualitative results in Sec. S-G.

S-A. Experiment Setup

Data Details. Building upon the work of previous studies [4], this research utilizes commonly used datasets such as Human3.6M [5], COCO [12], MPII [1], and MPI-INF-3DHP [15]. To ensure the quality of HardMo, we first discard ineligible samples using various filtering methods following [4]. First, we discard the keypoints with confidence less than 0.5. Then we discard samples with few 2D keypoints, unusual shapes, and unusual poses. Finally, we split the remaining dataset by subject. For HardMo-Foot P2, we choose jazz, Chinese dance, and ballet as the training set and samba, waltz, chacha, and tango as the testing set. For HardMo-Hand, training and testing are performed on the same motion classes, e.g., jazziness. It is split by the subject, to ensure no overlap. Furthermore, to ensure the quality of the testing set, we also manually check some samples.

Keypoints Details. We use 2D keypoints in COCO-WholeBody format [7, 19]. As shown in Fig. S-1, we use joints indexed from 1 to 23 as body and foot keypoints. For hand keypoints, we select every four joint indexed from 96 to 112 and from 117 to 133, e.g. 96, 100, and so on. Since we use the SMPL [13] models, we discard the face keypoints.

Implement Details. Following [2], we train HMR [9] using an HRNet-W48 backbone [3, 16, 18]. We utilize a batch size of 392 to train HMR and employ an AdamW [14] optimizer with a learning rate set at 1e-4, a weight decay set at 1e-4, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For different losses and data augmentation, we use the same settings as 4DHumans [4].

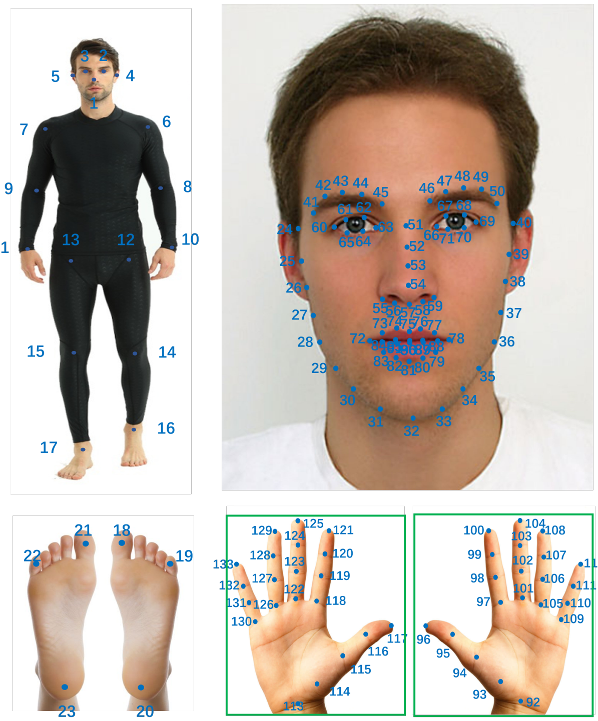

Figure S-1. **COCO-WholeBody Keypoints Format**. Because SMPL [13] is applied in this paper, we discard face keypoints. For body and foot keypoints, joints indexed from 1 to 23 are selected. For hand keypoints, we select every four joint indexed from 96 to 112 and from 117 to 133.

S-B. Metrics

For our evaluation, we use the metric as follows:

3D Metrics. Following previous work [4], we use MPJPE and PA-MPJPE for 3D evaluation. MPJPE (Mean Per Joint Position Error) is calculated as the mean L2 error between the predicted joints and ground-truth joints after aligning the root. PA-MPJPE (Procrustes analysis MPJPE) is MPJPE after the alignment of the predicted joints with ground-truth joints using the Procrustes Analysis.

2D Metrics. We follow [4], using the PCK as 2D evaluation metrics. PCK denotes Percentage of Correct Keypoints. A keypoint is deemed accurate if its Euclidean distance to the ground-truth is less than a threshold. We choose different thresholds (0.01 and 0.05 of image size).

Method	HardMo-Foot P1								HardMo-Foot P2							
	MPJPE↓		PA-MPJPE↓		PCK@0.01↑		PCK@0.05↑		MPJPE↓		PA-MPJPE↓		PCK@0.01↑		PCK@0.05↑	
	Body	Foot	Body	Foot	Body	Foot	Body	Foot	Body	Foot	Body	Foot	Body	Foot	Body	Foot
HardMo-HMR (w/o OPT)	38.9	58.3	23.4	16.2	0.57	0.43	**0.99**	0.98	42.1	58.1	25.6	21.1	0.57	0.51	0.98	0.98
HardMo-HMR (w/ OPT)	**23.6**	**34.9**	**15.4**	**10.8**	**0.71**	**0.58**	**0.99**	**0.99**	**27.5**	**39.5**	**17.7**	**13.2**	**0.67**	**0.60**	**0.99**	**0.99**

Table S-1. Ablation study of Foot-Hardcase optimaztion process. (1) P1 is intra-class, P2 is inter-class. (2) OPT denotes the optimized label.

Method	PCK@0.01		PCK@0.05	
	Body	Hand	Body	Hand
HardMo-HMR	0.445	0.246	0.975	0.961
HardMo-HMR (w/ F.S.O)	**0.616**	0.285	**0.987**	0.975
HardMo-HMR (w/ S.S.O)	0.490	**0.363**	0.974	**0.978**
HardMo-HMR (w/ B.S.O)	0.543	0.362	0.976	**0.978**

Table S-2. Ablation study of the Hand-Hardcase optimization process. F.S.O, S.S.O, and B.S.O represent first stage optimization, second stage optimization and both stage optimization, respectively.

Method	3DPW		LSP-Extended	
	PA-MPJPE↓	MPJPE↓	PCK@0.05↑	PCK@0.1 ↑
4DHumans^b	54.5	81.4	**0.54**	**0.86**
HardMo-4DHumans	**54.4**	**80.9**	0.82	0.84

Table S-3. Reconstruction error on 3DPW and LSP-Extended.

S-C. Evaluation on Automatic Annotation Pipeline

Ablation Study. To validate the effectiveness of our automatic annotation pipeline, we conduct ablation studies for pseudo labels optimization on HardMo-Foot and HardMo-Hand. We use HMR [9] trained on HardMo as our baseline, and the detailed results are shown in Table S-1 and S-2.

For foot-hardcase, we train the HMR [9], following two settings: (i) using the raw SMPL [13] annotations, denoted as HardMo-HMR (w/o OPT); (ii) using the SMPL annotations with optimization, denoted as HardMo-HMR (w/ OPT). As reported in Table S-1, the training with pseudo-label optimization significantly improves the performance of HMR [9] on HardMo-Foot P1, with body MPJPE decreasing by 15.3mm and foot MPJPE by 23.4mm compared to the model without the pseudo labels optimization. The above results prove the effectiveness of our optimization process for foot-hardcase and the importance of precise labels in solving hardcase. For hand-hardcase, the optimization process is divided into two stages. To verify the effects of both stages, we separately train the HMR model, following four settings: (i) using the raw SMPL annotations; (ii) F.S.O: using the SMPL annotations with the first stage optimization; (iii) S.S.O: using the SMPL annotations with the second stage optimization; (iv) B.S.O: using the SMPL annotations with both stages optimization. As reported in Table S-2, HardMo-HMR w/ F.S.O increases the PCK@0.01-body by 0.171 compared to the raw settings, proving the effectiveness of our first-stage optimization in improving the accuracy of body parts. HardMo-HMR w/ S.S.O increases the PCK@0.01-hand by 0.117 compared to the raw settings, demonstrating the crucial role of our second-stage optimization in correcting hand hardcase. Moreover, HMR w/ B.S.O increases the PCK@0.01-body by 0.053 to 0.543 compared to HardMo-HMR w/ S.S.O., and also almost no change in the PCK@0.01-hand. It proves that the pseudo-labels after both stages of optimization can improve hand precision while minimizing the impact on the accuracy of other body parts.

Method	PA-MPJPE↓	MPJPE↓
HardMo-4DHumans(w/o OPT)	41.5	64.0
HardMo-4DHumans(w OPT)	**40.7**	**62.5**

Table S-4. Reconstruction error on HardMo

Reason for not optimizing on HardMo-Normal. In preliminary experiments, we performed label optimization on subsets of HardMo-Normal. We found that the improvement is close to saturation when the subset reaches around 350k. This practice results in a marginal improvement of only 0.8 in PA-MPJPE, which is quite limited. Moreover, optimizing the whole HardMo-Normal requires about 6000 GPU hours. Considering the trade-off between overhead and improvement, we thus opt to optimize labels only on hardcase subsets.

Our decision was influenced by preliminary experiments on a 350k subset of the training set. Table S-4 shows that HardMo-4DHumans-350k (w OPT) slightly outperforms HardMo-4DHumans-350k (w/o OPT) in HardMo-test. Furthermore, optimizing the HardMo-Normal model requires 6000 GPU hours. Consequently, considering the minimal impact of optimization, we opted not to optimize the HardMo-Normal dataset.

S-D. Evaluation on 3DPW and LSPet

To validate the effectiveness of HardMo datasets, We also conduct experiments on commonly used benchmarks such as 3DPW or LSPet. The results are presented in Table S-3. Our HardMo-4DHumans is on par with or outperforms vanilla 4DHumans on both datasets. It indicates HardMo-4DHumans not only performs well on Hardcase scenes, but also performs well on common scenes.

S-E. Impact of Data-Scale

In this section, we validate how data scaling enhances performance. We train the HMR model [9] with instances of 5.8M, 2.9M, and 1.4M. As reported in Fig. S-2, the HMR trained with 5.8M instances achieves an MPJPE of 36.0mm, showing a 16.3% improvement over the HMR trained with 1.4M instances on the HardMo test set. These results further validate the effectiveness of data scale in boosting performance. With the advantage of such large training instances, HMR [9] can surpass the SOTA [4] on our benchmark.

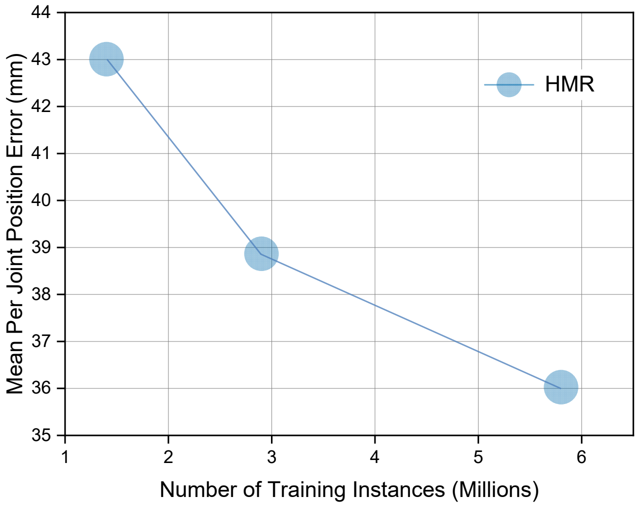

Figure S-2. **Data Scaling up**. As depicted, data scaling notably enhances the performance of HMR [9]. When the training instances are increased from 1.4M to 5.8M, there is a significant reduction in MPJPE, decreasing from 43.0mm to 36.0mm.

S-F. Collection Process

As previously mentioned in the main paper, existing datasets lack scenes of dance and martial arts, thus making it difficult for SOTA methods to perform well within these scenes. Additionally, all current methods struggle with hand-hardcase and foot-hardcase, where the resulting foot and hand poses often incorrectly resemble a T-pose. Building on the insights mentioned above, we establish two core principles for collecting the HardMo dataset: 1. Collect videos with a wide range of dance and martial arts types and diverse scenarios. 2. Place a special emphasis on the hand and foot hardcase. Following these rules, our data collection proceed as follows:

Initially, we identify 29 action categories for collection, including 15 dance styles and 14 types of martial arts. This categorization is based on existing dance datasets such as AIST++ [11, 17] and augmented by suggestions from LLM [6]. This category count is four times that of the AIST++ dataset. Drawing from the categories mentioned above, we gathered over 1,500 high-quality videos of dance and martial arts from the internet. To ensure the quality of the videos, we conduct a human check on the downloaded content, eliminating any that did not meet our quality standards. Diversity of character and scenarios is indispensable for enhancing model performance in dance and martial arts. Therefore, during our search process, we devise a set of prompts for both generic and specific action scenes, such as practice rooms, outdoors, stages, and competition venues. Additionally, to increase the proportion of foot-hardcase samples in the dataset, we collect more videos of ballet and jazz. After gathering these videos, we employ the YOLOv8 [8] tracking algorithm to automatically filter out leader and trailer segments without people and scenes with overly crowded characters. Following the above strategy, we gather a dataset named HardMo, consisting of over 7 million frames, labeled with 2D keypoints and SMPL annotations. Moreover, through selection and optimization, we establish two specialized subsets: a >500K HardMo-Foot dataset and a >400K HardMo-Hand dataset.

S-G. Additional Qualitative results

In the main paper, we have provided the visualization of annotation on HardMo, Comparisons on unusual poses, and Comparisons on Hardcase samples. In this section we provide additional qualitative results as follows:

Visualization of Annotation on HardMo. Figures S-3 and S-4 present the visualization of SMPL [13] annotations on HardMo-Hand and HardMo-Foot, respectively. As shown, our SMPL annotations are highly precise. Crucially, after our optimization process, we obtain annotations with accurate hand and foot poses.

Comparisons on unusual poses. In Fig. S-5, we present additional visualization results comparing HardMo-HMR with existing methods. As reported, HardMo-HMR surpasses ProHMR [10] by a notable margin and achieves performance on par with the state-of-the-art method 4DHumans [4].

Comparisons on Hardcase samples. In Fig. S-6, we provide additional qualitative results for tackling foot and hand hardcase problems. Here, we compare HardMo-4DHumans with existing methods. As reported, HardMo-4DHumans solve the inherent hardcase problems compared to the ProHMR [10] and 4DHumans [4]

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1

[2] Michael J Black, Priyanka Patel, Joachim Tesch, and Jin-

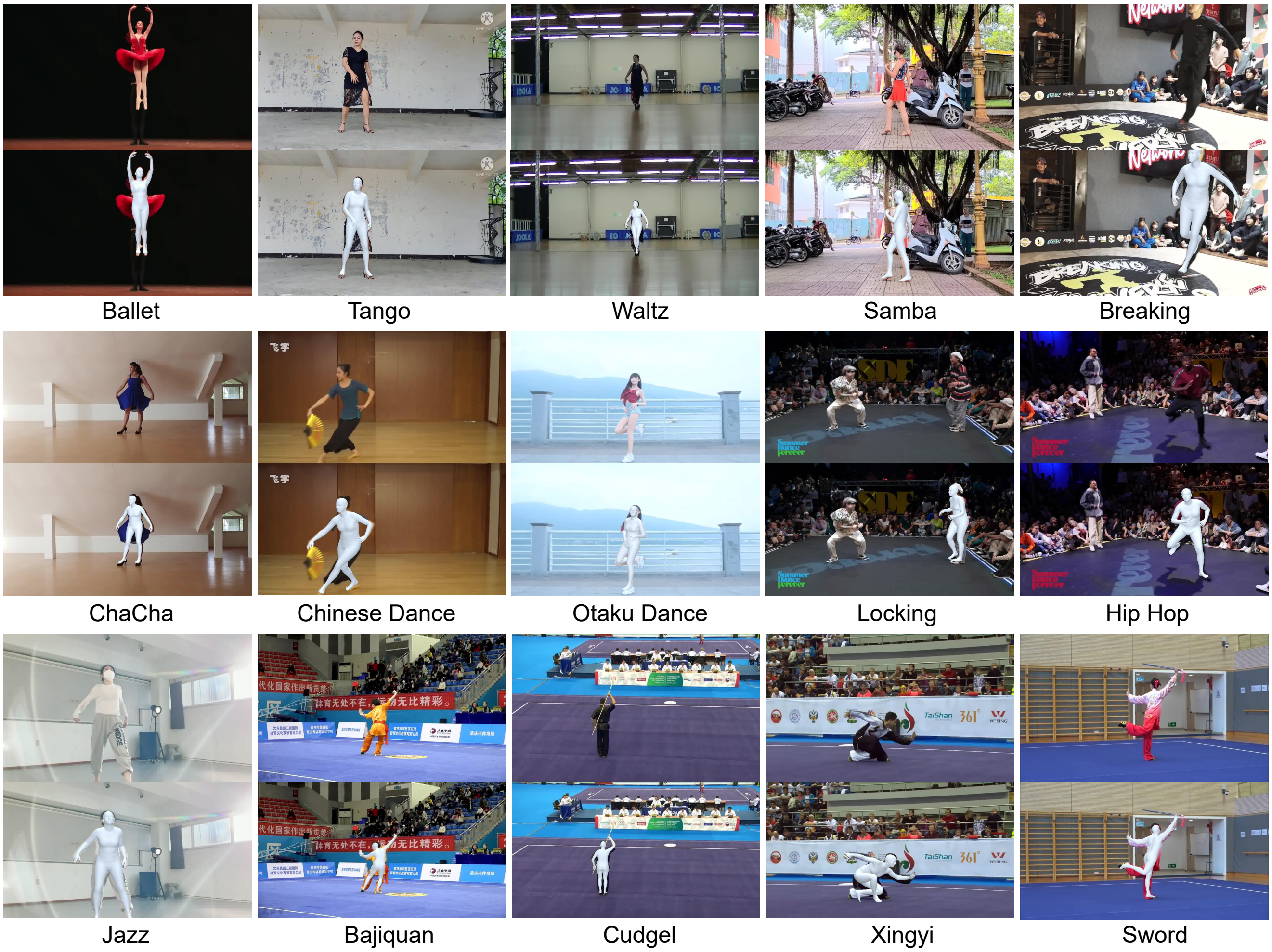

Figure S-3. SMPL [13] annotations for the 15 classes in HardMo-Hand. For each category, we display the input image (top) and the annotated image (bottom).

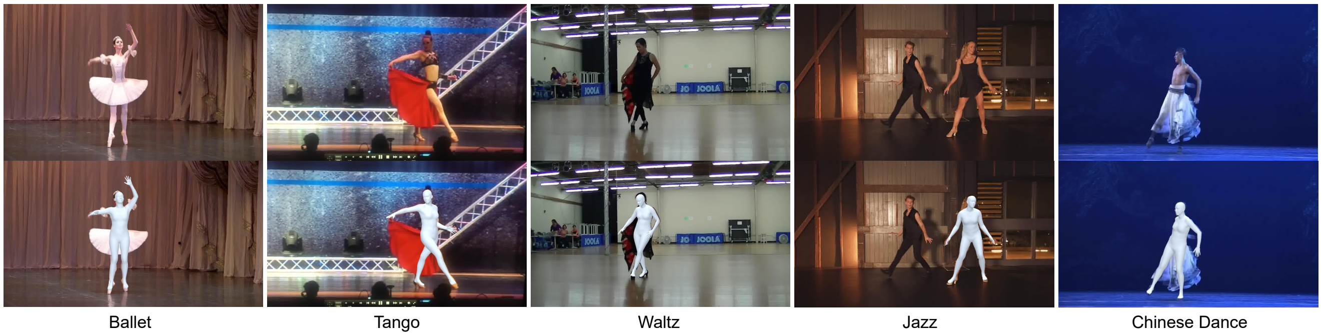

Figure S-4. SMPL [13] annotations for the 5 classes in HardMo-Foot. For each category, we display the input image (top) and the annotated image (bottom).

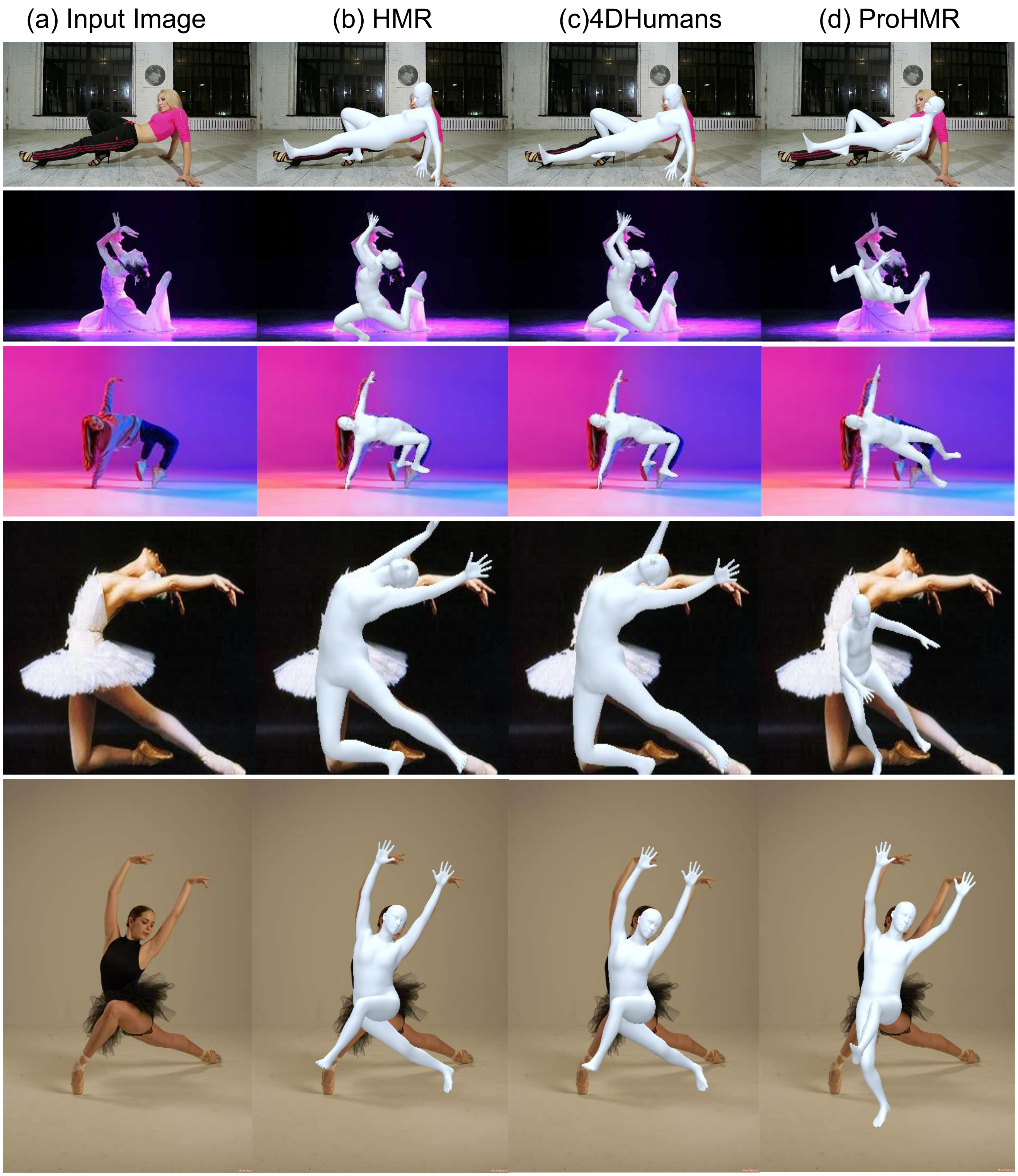

Figure S-5. Comparisions of HardMo-HMR and existing methods on unusual poses

.

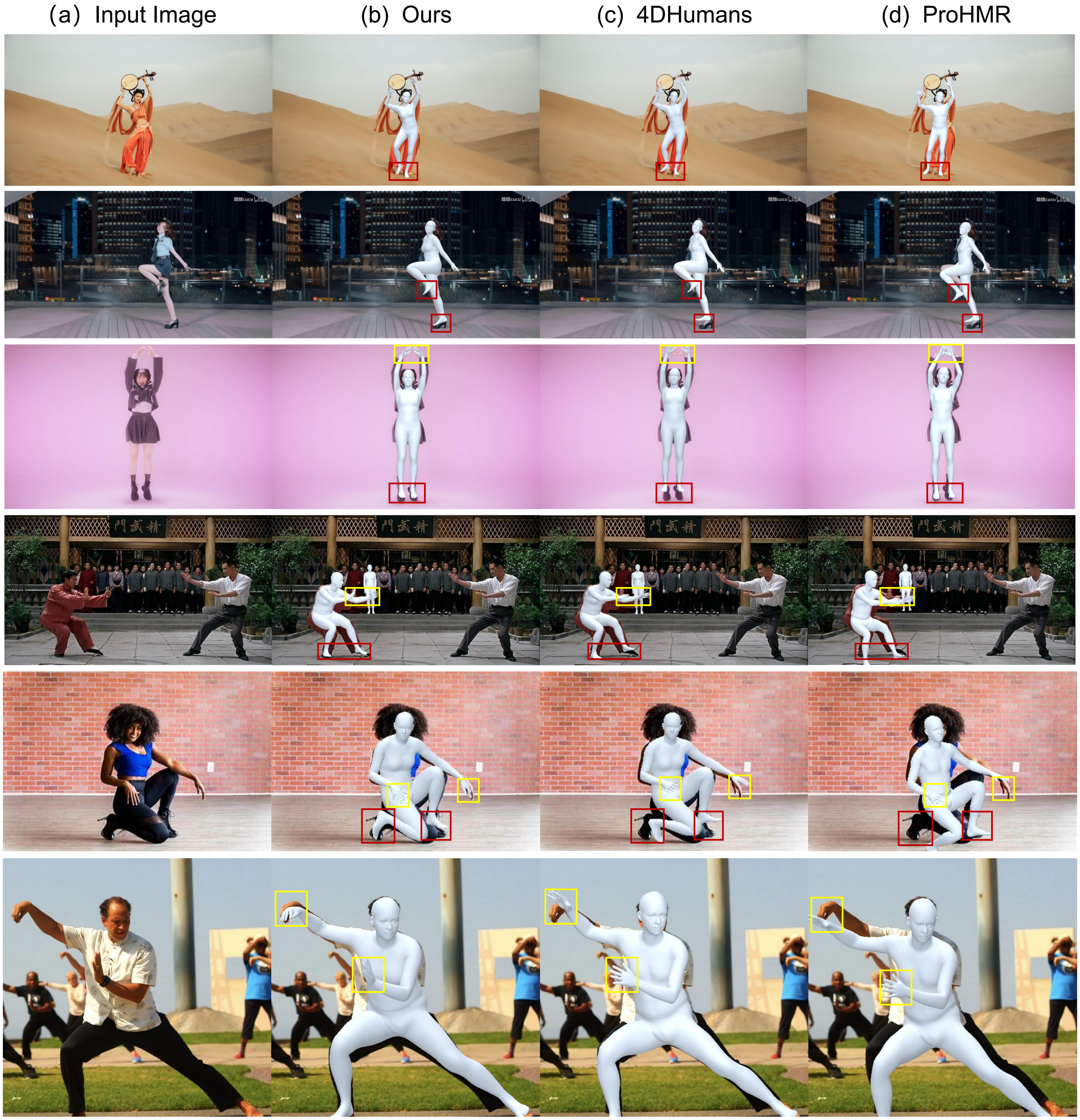

Figure S-6. Comparisions of HardMo-4DHumans and existing methods on hardcase samples

.

long Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 1

[3] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 1

[4] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 3

[5] 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. 1

[6] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 2023. 3

[7] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[8] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 2023. 3

[9] 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. 1, 2, 3

[10] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11605–11614, 2021. 3

[11] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 3

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[13] 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)*, 34(6):248:1–248:16, 2015. 1, 2, 3, 4

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[15] 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 *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 1

[16] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1

[17] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, 2019. 3

[18] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. 1

[19] Lumin Xu, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Zoomnas: Searching for whole-body human pose estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1