HardMo: A Large-Scale Hardcase Dataset for Motion Capture

Jiaqi Liao[1*], Chuanchen Luo[2*], Yinuo Du[1], Yuxi Wang[2,3],
Xucheng Yin[5], Man Zhang[1,6], Zhaoxiang Zhang[2,3,4✉], Junran Peng[5✉]
[1] Beijing University of Posts and Telecommunications [2] Institute of Automation, Chinese Academy of Sciences
[3] Centre for Artificial Intelligence and Robotics, HKISI, CAS [4] University of Chinese Academy of Sciences
[5] University of Science and Technology Beijing [6] Qinghai University of Science and Technology
{liaojiaqi, duyinuo99, zhangman}@bupt.edu.cn, zhaoxiang.zhang@ia.ac.cn
{chuanchenluo,yuxiwang93,jrpeng4ever}@gmail.com, xuchengyin@ustb.edu.cn

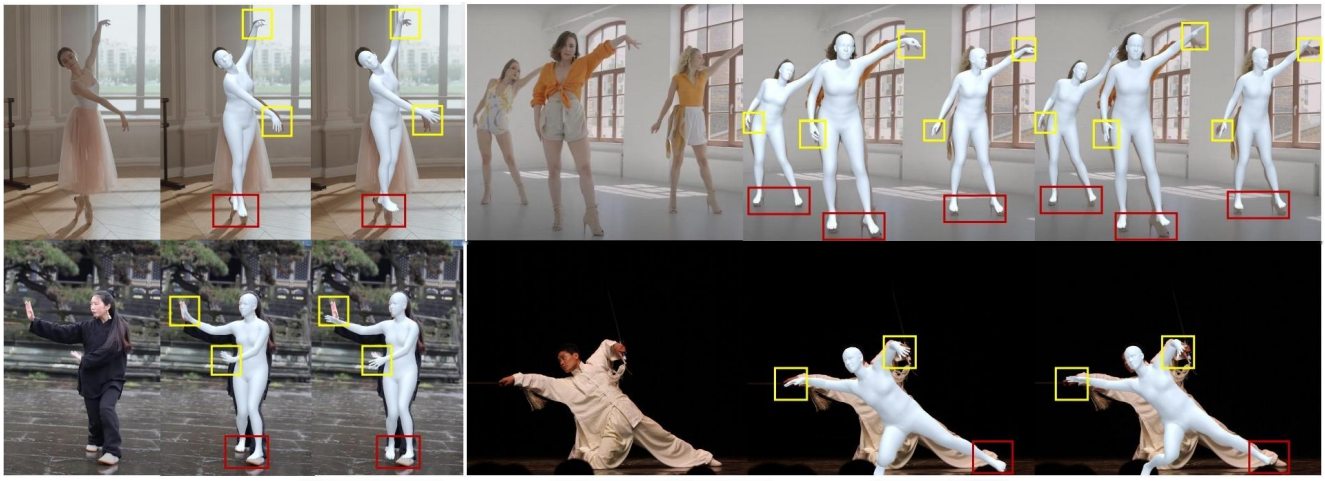

Figure 1. Comparisons between vanilla 4DHumans [10] and our HardMo-4DHumans. The teaser comprises four sub-figures. Each sub-figure, from left to right, corresponds to (a) the input image, (b) the prediction of HardMo-4DHumans, and (c) the prediction of vanilla 4DHumans. In comparison, our HardMo-4DHumans is superior in the alignment of hand and foot posture.

Abstract

Recent years have witnessed rapid progress in monocular human mesh recovery. Despite their impressive performance on public benchmarks, existing methods are vulnerable to unusual poses, which prevents them from deploying to challenging scenarios such as dance and martial arts. This issue is mainly attributed to the domain gap induced by the data scarcity in relevant cases. Most existing datasets are captured in constrained scenarios and lack samples of such complex movements. For this reason, we propose a data collection pipeline comprising automatic crawling, precise annotation, and hardcase mining. Based on this pipeline, we establish a large dataset in a short time. The dataset, named HardMo, contains 7M images along with precise annotations covering 15 categories of dance and 14 categories of martial arts. Empirically, we find that the prediction failure in dance and martial arts is mainly characterized by the misalignment of hand-wrist and foot-ankle. To dig deeper into the two hardcases, we leverage the proposed automatic pipeline to filter collected data and construct two subsets named HardMo-Hand and HardMo-Foot. Extensive experiments demonstrate the effectiveness of the annotation pipeline and the data-driven solution to failure cases. Specifically, after being trained on HardMo, HMR, an early pioneering method, can even outperform the current state of the art, 4DHumans, on our benchmarks. Dataset will be publicly available at https://ljqnb.github.io/HardMo.github.io.

*Equal contribution.

Type	Dataset	#Frames	#Scenes	#Subjects	Scene Type	Dance&Martial Arts Type	Hardcase	Annotation Type
Rendered Dataset	AGORA [29]	17K	>350	4,240	Daily	-	-	SMPL-X
	BEDLAM [4]	380K	103	-	Daily	-	-	SMPL-X
Marker/Sensor-based MoCap	Human3.6M [13]	3.6M	1	11	Daily	-	-	SMPL
	3DPW [38]	> 51K	60	7	Daily	-	-	SMPL
Marker-less Multi-view MoCap	MPI-INF-3DHP [28]	>1.3M	1	8	Daily	-	-	SMPL
	AIST++ [23]	10.1M	1	30	Dance	10	-	SMPL
Pseudo-3D Labels	MSCOCO [25]	38K	-	-	Daily	-	-	SMPL
	MPII [1]	24,920	3,913	>40k	Daily	-	-	SMPL
	HardMo(Ours)	>7.0M	>350	933	Dance&Martial Arts	29	-	SMPL
	HardMo-Hand(Ours)	>400K	>320	848	Dance&Martial Arts	29	✓	SMPL
	HardMo-Foot(Ours)	>500K	>180	102	Dance	7	✓	SMPL

Table 1. Comparison with related datasets.

1. Introduction

Monocular motion capture aims to recover human skeletal motions from single-view videos. As a pivotal component of computer animation, this technique is primarily employed to endow virtual characters with authentic motion, especially in the context of dance and martial arts. For example, most dance and martial arts in games (*e.g. Genshin Impact*) and movies (*e.g. Avatar*) are realized through the motion capture technique. Fueled by deep learning techniques, recent years have witnessed remarkable progress in monocular motion capture. Since the proposal of HMR [18], a series of methods [18, 22, 24, 43] have emerged and achieved great breakthroughs on public benchmarks. Specifically, the MPJPE metric on Human3.6M [13] has seen a significant decrease from 88.0 mm to 47.1 mm.

Despite the promising performance on benchmarks, these methods underperform in dance and martial arts scenarios. Most existing methods are vulnerable to unusual poses in dance and martial arts scenes, revealing the vast gap between research and practical application. From a data-driven perspective, the crux lies in the *domain gap* between existing motion datasets and real-world scenarios. In stark contrast to daily actions, dance and martial arts are characterized by rapid and tension-filled skeletal movements. As shown in Table 1, such movements rarely appear in commonly used datasets such as Human3.6M [13], MPI-INF-3DHP [28], and COCO [25]. As a result, models trained on these standard datasets struggle to effectively handle dance and martial arts.

Recently, 4DHumans [10] additionally employed the InstaVariety [19] dataset for training. This practice leads to a promising improvement in handling unusual poses which demonstrates the effectiveness of the data-driven paradigm. Nevertheless, the Instavariety [19] dataset lacks diversity. It contains a limited number of dances and excludes martial arts. Thus, this dataset cannot fully bridge the domain gap. Moreover, as shown in Fig. 1, 4DHumans [10] trained on Instavariety [19] still suffers from prediction error in two cases, *i.e. foot-hardcase* (incorrect foot-ankle posture) and *hand-hardcase* (incorrect hand-wrist posture), where hands and feet tend to be recovered to the rest pose, respectively. From a data-driven perspective, we identify three limitations that may cause the issue: **(1) Limited Diversity in Hand and Foot Postures.** Synthetic datasets have an advantage in precise SMPL [26] annotations. But they mainly focus on daily actions and lack diversity in hand and foot postures. **(2) Incorrect SMPL Annotations.** Some real-world datasets, such as Human3.6M [13], have incorrect hand and foot annotations. The two body parts are annotated as a nearly rest pose. **(3) Keypoint Annotations Lack Hand and Foot.** Keypoints annotations of existing datasets are either in COCO format or in Human3.6M format. The two formats only contain body keypoints and lack hand and foot keypoints. Using such annotations to train models would lead to a lack of constraints on hand and foot joints.

According to the analysis mentioned above, how to collect relevant data with precise annotations is crucial to handle domain gap and inherent hardcase issues. Considering the highly technical and artistic demands in dance and martial arts, collecting data using marked or markerless capture methods would be prohibitively expensive and might not be sufficiently comprehensive. Therefore, we propose an automatic pipeline that leverages online videos. First, we gather ample dance and martial arts videos from the Internet. Second, we employ RTM-pose [9, 14] and 4DHumans [10] to estimate 2D keypoints and raw SMPL parameters, respectively. Third, we propose an angle-based hardcase mining algorithm to identify foot-hardcase and hand-hardcase samples. Lastly, we further optimize the SMPL parameters of these hardcases, since the precise annotation is crucial for handling the issue caused by data bias. Such a pipeline can accurately annotate 1 million images within three days, showcasing its efficiency, precision, and scalability.

Based on this pipeline, we introduce HardMo, a large-scale hardcase dataset for monocular motion capture. HardMo contains over 7 million images extracted from

1,500 sequences spanning 15 categories of dance and 14 categories of martial arts. Each image is accompanied by 2D keypoints and SMPL [26] annotations. Furthermore, to specifically tackle foot-hardcase and hand-hardcase issues, we curate additional HardMo-Foot and HardMo-Hand subsets which contain over 500k and 400k samples of corresponding hardcases, respectively.

In summary, our contributions are threefold:

- We develop an efficient and scalable pipeline for automatic annotation and hardcase mining. This system offers a potent solution to the data scarcity issue in the motion domain.
- HardMo bridges the domain gap, containing 7 million images across over 300 different scenarios. As subsets of HardMo, HardMo-Hand and HardMo-Foot, the first of their kind, focus on solving the inherent hardcase issues.
- Extensive experiments demonstrate the effectiveness of HardMo in addressing the domain gap and inherent hardcase issues.

2. Related Work

2.1. Human Mesh Recovery

The techniques for human mesh recovery can generally be categorized into two main approaches: optimization-based [5, 12, 30, 34, 36, 41, 42] and regression-based [3, 8, 10, 18–22, 24, 27, 31, 43, 44]. The predominant method in the optimization approach iteratively optimizes by fitting the pose and shape parameters of SMPL [26] based on 2D keypoints. The Optimization method typically yields highly accurate results when the quality of the 2D keypoints is reliable. However, when faced with the occlusions, and unusual movements in dances and martial arts scenes, the results of optimization can be unsatisfactory. With the rise of deep learning, regression-based methods have increasingly become mainstream. Starting with HMR [18], many techniques have been developed to improve upon its foundation. For example, SPIN [21] utilizes the regression results of HMR as the initial pose for SMPLify [5]. PyMAF [43, 44] introduces a mesh alignment module to correct poorly performing regression results, while CLIFF [24] incorporates additional bounding box input and calculates the 2D reprojection loss on the full image. Nonetheless, despite their excellent performance on benchmarks such as 3DPW [38] and Human3.6m [13], these methods often fall short in real-world scenarios, particularly in dance and martial arts. However, after training additionally on the Instavariety [19] dataset, 4DHumans [10] has shown significant advancements and can capture some unusual poses well. Nevertheless, all the methods above still struggle with two inherent hardcase issues.

2.2. Human Body Pose Datasets

As shown in Table 1, The available datasets can generally be grouped into four types: rendered datasets [4, 6, 29], marker-based mocap datasets [7, 13, 35, 38], markerless mocap datasets [16, 23, 28, 32, 39], pseudo-label datasets [1, 2, 25, 45]. (1) Rendered datasets, such as AGORA [29] and BEDLAM [4], offer the most standard SMPL [26] labels. However, these datasets lack realism and don't include dance or martial arts scenes. (2) In contrast, motion capture datasets like Human3.6M [13] are limited to a singular scenario, with a very basic range of movements. Although there are outdoor motion capture datasets like 3DPW [38], they still don't include dance or marital arts scenes. (3) To enrich the diversity of actions, markless motion capture datasets like AIST++ [23, 37] have emerged. They capture different views of 2D images and joints and then employ triangulation techniques to get 3D joints, subsequently fitting the SMPL model to these 3D joints to produce pseudo labels. While AIST++ focuses on dance, its dataset lacks variety in scenarios and dance types. Critically the poses of hand and foot in their dataset are inaccurately represented, resembling a T-pose. (4) Given the known richness and diversity of 2D pose datasets in subjects, poses, and scenes, some methods apply pseudo labels to these 2D datasets. However, as mentioned in Sec. 1, the hand and foot of their pseudo SMPL labels often resemble a T-pose, which causes the hardcase problems.

3. HardMo Dataset

The HardMo dataset is a large-scale hardcase dataset for motion capture in dance and martial arts scenes. It is distinguished from existing datasets by the emphasis on hardcases of two scenarios close to practical deployment. The HardMo dataset contains over 7 million images with precise 2D keypoints and 3D SMPL [26] annotations. These images are collected from 1,500 sequences covering 15 categories of dance and 14 categories of martial arts. Such a dataset bridges the gap between current human mesh recovery methods and real-world applications effectively. In dance and martial arts scenes, we observe that the misalignment of hands and feet appears frequently. To deal with the two hardcases, we additionally curate a HardMo-Foot dataset and a HardMo-Hand dataset. The two datasets contain over 500K and 400K samples with unusual foot and hand pose, respectively. To ensure the efficacy of both datasets, we perform further optimization on the SMPL annotations of these samples. For better understanding, we show some examples of our dataset in Fig. 2.

4. Automatic Annotation Pipeline

To process massive raw videos efficiently, we develop a pipeline for automatic annotation, hardcase mining, and la-

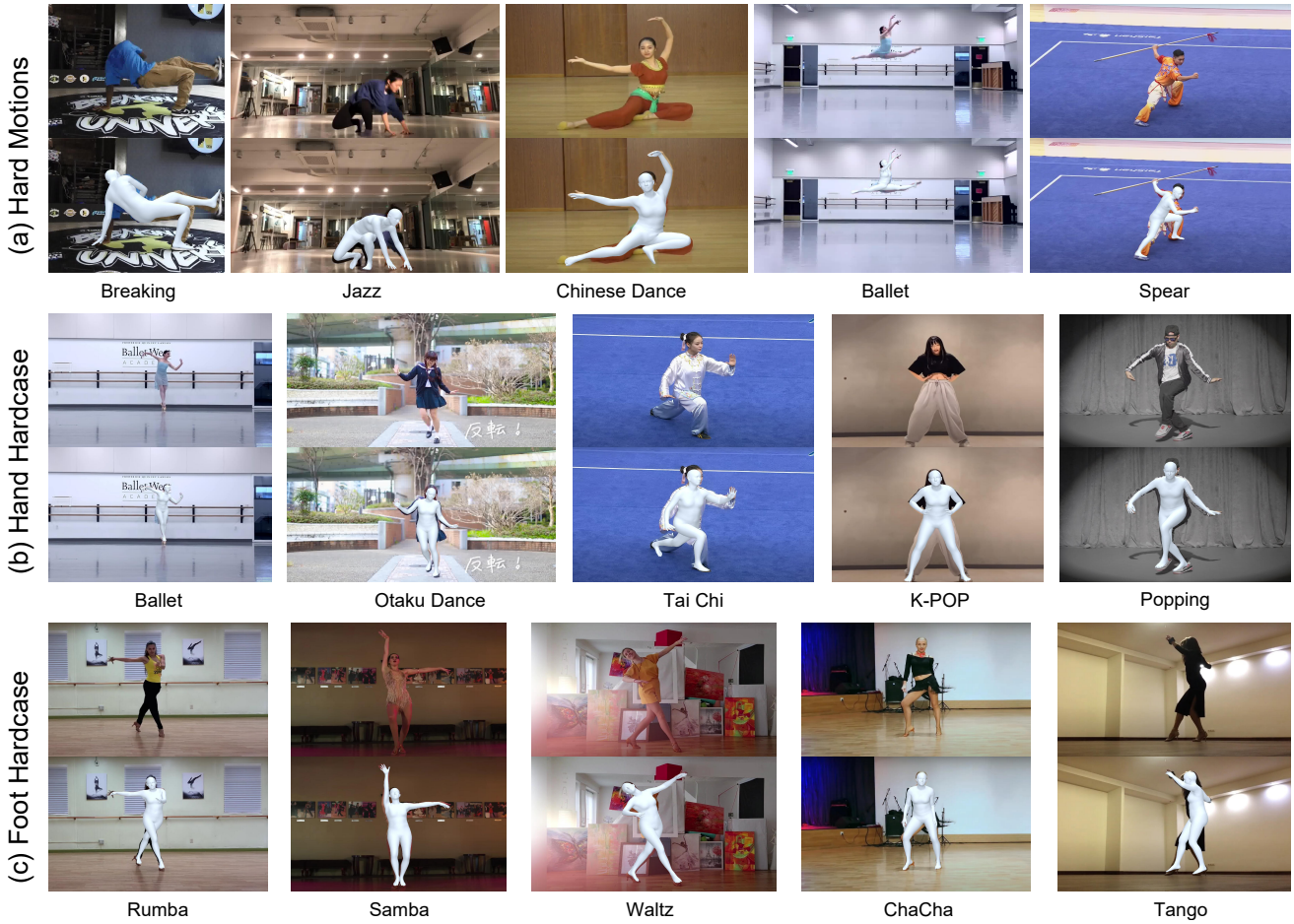

Figure 2. Overview of HardMo. It contains: (a) challenging and artistic motions from diverse types of dance and martial arts, (b) Hand-hardcase with precise annotations, and (c) Foot-hardcase with precise annotations.

bel optimization. The pipeline has two stages: First, the Normal Annotation stage annotates the HardMo dataset with pseudo labels predicted by off-the-shelf regressors. Then, the Hardcase Annotation stage filters out hardcases and refines their labels to obtain the HardMo-Foot and HardMo-Hand datasets. For more implementation details of the pipeline, please refer to the supplementary material.

4.1. Normal Annotation

Given a video sequence, we first employ YOLOv8 [15, 33] to detect bounding boxes of persons. Then, the person images are cropped accordingly for subsequent annotation. In terms of 2D keypoint annotation, we apply RTM [9, 14] to estimate the whole-body keypoints $\mathbf{K}^{\text{2D}}$ from the cropped image including body keypoints $\mathbf{K}^{\text{2D}}_{\text{body}} \in \mathbb{R}^{23\times 2}$ and hand keypoints $\mathbf{K}^{\text{2D}}_{\text{hand}} \in \mathbb{R}^{42\times 2}$. We discard the facial keypoints since SMPL does not cover facial expressions. To ensure the reliability of 2D keypoint annotations, we exclude samples with an average keypoint confidence below 0.5. As for 3D annotation, we predict the SMPL [26] parameters Θ using the state-of-the-art body mesh recovery method, 4D-Humans [10].

4.2. Hardcase Annotation

In preliminary experiments, SMPL annotations failed to meet our requirement for quality in some cases. We empirically find that the failures mainly derive from extreme hand and foot posture. To delve into the effect of these hardcases, we design an algorithm to collect such samples automatically. As mentioned above, these hardcases suffer from imprecise 3D SMPL annotations. To correct their annotations, an intuitive solution is to employ some optimization methods like SMPLify [5] or EFT [17]. However, these methods damage the annotation precision of other parts and may lead to implausible poses. For this reason, we instead devise a learning-based method to refine the SMPL annotations.

Hardcase Mining. To mine specific hardcase samples, we propose an angle-based mining algorithm. As mentioned in Sec. 1, the existing methods tend to recover meshes where

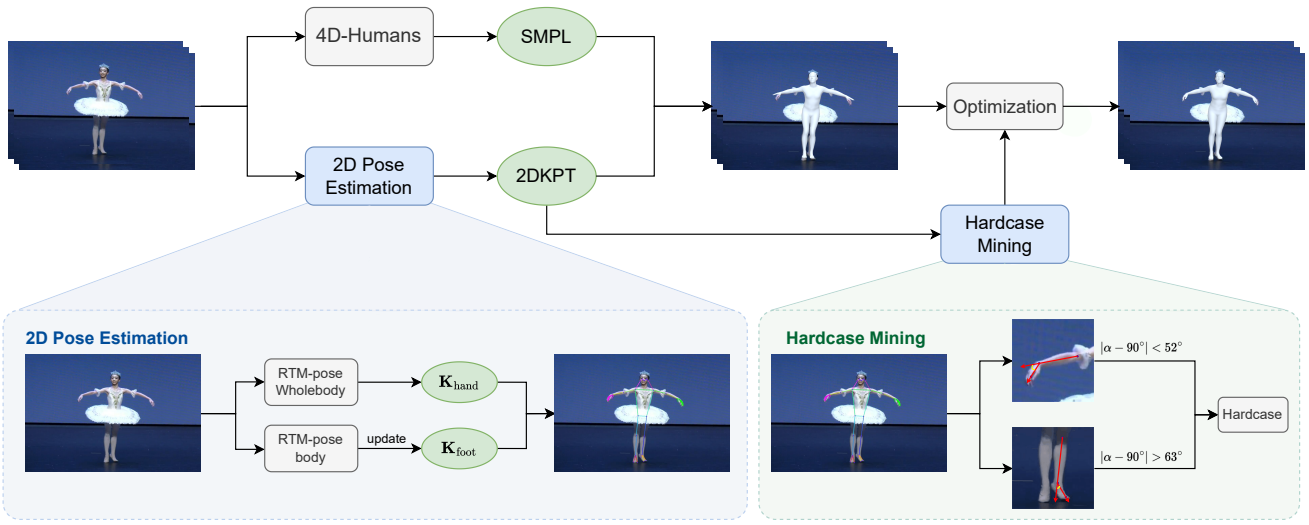

Figure 3. Overview of the automatic annotation pipeline. It contains 2D pose estimation, angle-based hardcase mining and hardcase-based optimization.

the hand and foot are close to T-pose. In this case, the ankle angle usually approaches $90°$, and the wrist angle tends towards $180°$. Based on these insights, we design our angle-based Hardcase mining algorithm as follows: For the foot, we can approximate the coordinate of mid-toe joint as:

$$\mathbf{K}_{\text{midtoe}} = \frac{\mathbf{K}_{\text{bigtoe}} + \mathbf{K}_{\text{smalltoe}}}{2} \tag{1}$$

According to Equation 1, we define the vectors for the foot and leg as shown in Fig. 3:

$$\mathbf{V}_{\text{foot}} = \mathbf{K}_{\text{midtoe}} - \mathbf{K}_{\text{ankle}}, \mathbf{V}_{\text{leg}} = \mathbf{K}_{\text{ankle}} - \mathbf{K}_{\text{knee}} \tag{2}$$

Hence, we can calculate the ankle rotation using the value of $\alpha_{\text{ankle\_rotation}}$ as determined by Equation 3.

$$\alpha_{\text{ankle\_rotation}} = \arccos\left(\frac{\mathbf{V}_{\text{foot}} \cdot \mathbf{V}_{\text{leg}}}{\|\mathbf{V}_{\text{foot}}\|\|\mathbf{V}_{\text{leg}}\|}\right) \tag{3}$$

We then mine the hardcase samples based on the following criteria: 1. The confidence for each of the foot keypoints must exceed 0.65, ensuring the reliability of the 2D keypoints. 2. $|\alpha_{\text{ankle\_rotation}} - 90°| > 63°$, impling that the angular deviation of the ankle from the T-pose should exceed $63°$.

As for the hand, since our goal is to resolve the rotation issue of the wrist joint, we do not consider the precise finger postures. Therefore, during the mining and optimization of the hand-hardcase, we have selected 10 keypoints of the hand part. Then we can approximate the coordinate of hand as:

$$\begin{aligned}\mathbf{K}_{\text{hand}} = \frac{1}{5}(&\mathbf{K}_{\text{thumb}} + \mathbf{K}_{\text{fore\_finger}} + \mathbf{K}_{\text{middle\_finger}} \\ &+ \mathbf{K}_{\text{ring\_finger}} + \mathbf{K}_{\text{pinky\_finger}})\end{aligned} \tag{4}$$

Based on the keypoints defined in Equation 4, we define the vectors for the hand and arm as shown in Fig. 3:

$$\mathbf{V}_{\text{hand}} = \mathbf{K}_{\text{hand}} - \mathbf{K}_{\text{wrist}}, \mathbf{V}_{\text{arm}} = \mathbf{K}_{\text{wrist}} - \mathbf{K}_{\text{elbow}} \tag{5}$$

Based on Equation 5, we can calculate the wrist rotation as follows.

$$\alpha_{\text{wrist\_rotation}} = \arccos\left(\frac{\mathbf{V}_{\text{hand}} \cdot \mathbf{V}_{\text{arm}}}{\|\mathbf{V}_{\text{hand}}\|\|\mathbf{V}_{\text{arm}}\|}\right) \tag{6}$$

Then, we select hand-hardcase samples based on the following criteria: 1. The confidence for each of the 10 hand keypoints must exceed 0.7 to ensure the reliability of the 2D keypoints 2. $|\alpha_{\text{wrist\_rotation}} - 90°| < 52°$. Due to page limits, we leave more details about hardcase mining in supplementary details.

Foot-hardcase Optimization. To correct the behavior of the model in hardcases, we need to obtain precise pseudo labels of the filtered hardcase samples. Although the classic SMPLify [5] method can accurately fit the SMPL [26] parameter Θ to 2D keypoints, it tends to generate physically implausible pseudo labels. To solve the issues, the EFT method [17] employ the trained HMR model as a optimization prior. However, Due to the lack of specific priors for these hardcases, it performs poorly on optimizing hardcase samples. Therefore, we propose our methods. With the following loss denoted by Equation 7, the model is trained on the hardcase dataset. Thus it can learn an underlying hardcase pose prior via a data-driven paradigm. Then we use the trained model to test on the hardcase dataset, thus obtaining our precise pseudo labels.

$$\begin{aligned}L_{\text{joint}} =&\lambda_{\text{body}}\|\hat{\mathbf{K}}^{\text{2D}}_{\text{body}} - \bar{\mathbf{K}}^{\text{2D}}_{\text{body}}\|_1 + \lambda_{\text{foot}}\|\hat{\mathbf{K}}^{\text{2D}}_{\text{foot}} - \bar{\mathbf{K}}^{\text{2D}}_{\text{foot}}\|_1 \\ &+ \lambda_{\text{smpl}}\|\hat{\Theta} - \Theta\|_1.\end{aligned} \tag{7}$$

Method	MPJPE↓	PA-MPJPE↓	PCK@0.01↑	PCK@0.05↑
ProHMR [22]	113.8	74.2	0.10	0.73
CLIFF [24]	93.3	56.2	0.29	0.90
HMR [18]‡	61.1	42.1	0.40	0.96
HardMo-HMR†	46.0	31.5	0.47	0.97
HardMo-HMR	**36.0**	**25.0**	**0.56**	**0.98**
4DHumans^a [10]	83.1	52.7	0.16	0.92
4DHumans^b [10]	36.6	23.0	0.48	**0.98**
HardMo-4DHumans†	29.9	20.2	0.55	**0.98**
HardMo-4DHumans	**26.0**	**18.0**	**0.60**	**0.98**

Table 2. Human mesh recovery accuracy on the HardMo dataset. †: trained with mix of HardMo and commonly used datasets. ‡: trained with InstaVariety [19]. 4DHumans^a: HMR 2.0a from [10] which is trained on commonly used datasets. 4DHumans^b: HMR 2.0b from [10] which is trained additionlly on Instavariety [19], AVA [11] and AI Challenger [40] .

Here, $\bar{\mathbf{K}}^{\text{2D}}_{\text{body}}$ is the 2D keypoints pseudo labels of body, and $\bar{\mathbf{K}}^{\text{2D}}_{\text{foot}}$ indicates 2D keypoints pseudo labels of the foot. $\hat{\mathbf{K}}^{\text{2D}}_{\text{body}}$ and $\hat{\mathbf{K}}^{\text{2D}}_{\text{foot}}$ represent 2D keypoints of body and foot projected by SMPL [26] mesh. $\bar{\Theta} = (\bar{\theta}, \bar{\beta})$ is the pseudo label. $\bar{\theta}$ is the pose parameter of the SMPL model, and $\bar{\beta}$ is the shape parameter. $\Theta = (\theta, \beta)$ is the SMPL parameter predicted by the model. The use of raw SMPL pseudo labels as a regularization is important. We find that using only 2D keypoints as weak supervision can damage the prior knowledge learned by the model, consequently undermining the physical plausibility of results generated by the model.

Hand-hardcase Optimization. The hand exhibits greater flexibility compared to the foot. So employing the same optimization strategy as with foot-hardcase would make it challenging for the model to learn a reasonable solution. However, utilizing SMPLify [5] can well annotate hardcase parts but could potentially compromise other parts of the body. Therefore, we propose a two-stage optimization method: In the first stage, we follow the optimization process outlined in Sec. 4.2, focusing on the body and foot parts. In the second stage, we introduce a hand-based optimization approach: for each sample, we optimize the following loss parameters:

$$L_{\text{joint}} = \lambda_{\text{2D}} \|\hat{\mathbf{K}}^{\text{2D}}_{\text{hand}} - \bar{\mathbf{K}}^{\text{2D}}_{\text{hand}}\|_1 + \lambda_\theta \|\hat{\theta_{\mathbf{0:19}}} - \theta_{\mathbf{0:19}}\|_1. \quad (8)$$

Here, $\mathbf{K}^{\text{2D}}_{\text{hand}}$ represents 2D keypoints of the hand, $\theta_{0:19}$ denotes the SMPL [26] pose parameters that exclude the hand. $\hat{}$ denotes the predicted object, and $\bar{}$ denotes pseudo-labels. By applying this term, we impose a regularization loss on the body part to prevent the hand optimization from compromising the physical plausibility of other body parts. Through this two-stage optimization process, we not only ensure the physical plausibility of the body part but also annotate hardcase-part well.

5. Experiments

To validate the impact of the proposed dataset HardMo, we initially conduct experiments to evaluate existing state-of-the-art approaches. Subsequently, we introduce two innovative benchmarks, *i.e. HardMo-Foot* and *HardMo-Hand*, to investigate the efficacy of the collected data in addressing hardcase issues. Due to the page limit, we leave the validation of the effectiveness of the automatic annotation pipeline and the description of the metrics in the supplementary details.

5.1. Impact of HardMo

To ensure the quality of dataset, we first discard ineligible samples using various filtering methods. The remaining dataset is divided into training set (80%) and testing set (20%). For the test data, we use the optimization approach in Sec. 4.2 to ensure the accuracy of labels. For comprehensive evaluation, we design two evluation settings. The first one evaluates the performance of previous models trained on the commonly used datasets. The second one is reproducing existing methods on our HardMo dataset, including HMR [18] and 4DHumans following [10].

Results and Analysis. In Table 2, we present the comparison results on the HardMo benchmark. HardMo-HMR, trained solely on HardMo, achieving an MPJPE of 36.0 mm performance, significantly outperforms HMR [18] that is trained with InstaVariety [19] with 61.1 mm. Moreover, it is 3.2× better than ProHMR [22]'s 113.8 mm and even surpasses the SOTA method, 4DHumans [10] with 36.6 mm. These results show the effectiveness of HardMo in solving domain gap issues. Furthermore, HardMo-HMR†, trained on a mixed dataset about HardMo and commonly used datasets, shows a 10.0mm decrease in MPJPE compared to HardMo-HMR, which highlights the inadequacy of the commonly used datasets. Similar results are observed in the HardMo-4DHumans. What's more HMR‡ trained solely on the InstaVariety [19] dataset, shows a 25.1mm decrease in MPJPE compared to HardMo-HMR. This result demonstrates that compared to InstaVariety, HardMo is closer to real-world scenarios and more effective in mitigating the domain gap.

Visible Results. In Fig. 4, we present visualization results comparing HardMo-HMR with existing methods. The results reveal that the proposed HardMo-HMR excels in handling challenging poses, outperforming ProHMR [22], and even surpassing the state-of-the-art 4DHumans [10].

5.2. Hardcase Benchmarks

To show the effectiveness of our hardcase datasets on solving hardcase problems, we establish two benchmarks based on HardMo-Foot and HardMo-Hand datasets, respectively.

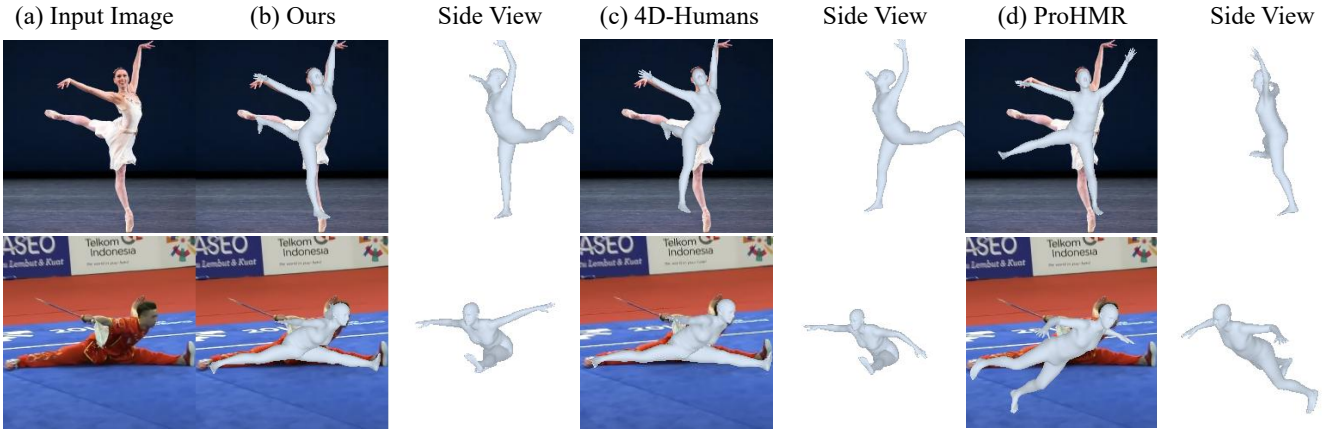

Figure 4. Qualitative comparison between HardMo-HMR and other leading methods. As shown in the figure, HardMo-HMR is superior in handling challenging poses.

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
ProHMR [22]	117.0	213.6	73.2	53.1	0.13	0.02	0.77	0.50	90.0	131.9	54.4	41.5	0.14	0.03	0.82	0.64
CLIFF [24]	93.5	169.4	49.5	42.0	0.31	0.14	0.91	0.82	73.1	120.4	39.5	32.0	0.42	0.22	0.97	0.92
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**
4DHumans^a [10]	86.7	143.7	47.2	41.8	0.16	0.08	0.95	0.79	77.1	100.0	40.1	33.3	0.22	0.14	0.98	0.92
4DHumans^b [10]	40.0	88.1	21.1	26.5	0.59	0.14	**0.99**	0.86	34.9	77.0	19.0	25.4	0.64	0.16	**0.99**	0.92
HardMo-4DHumans† (w/ OPT)	20.9	29.4	13.5	**9.0**	0.68	0.50	**0.99**	**0.99**	24.4	**34.5**	15.5	11.0	0.68	0.59	**0.99**	**0.99**
HardMo-4DHumans (w/ OPT)	**19.8**	**29.3**	**13.0**	**9.0**	**0.70**	**0.53**	**0.99**	**0.99**	**23.3**	34.8	**14.6**	**10.8**	**0.71**	**0.60**	**0.99**	**0.99**

Table 3. Reconstruction error on HardMo-Foot. P1 is intra-class evaluation, P2 is inter-class evaluation. †: trained on the mixture of HardMo-Foot and HardMo-Hand. OPT: optimized label. 4DHumans^a: HMR 2.0a from [10]. 4DHumans^b: HMR 2.0b from [10].

Besides, we also train 4DHuman [10] jointly on a mixture dataset of HardMo-Foot and HardMo-Hand.

5.2.1 Benchmark on HardMo-Foot

We conduct evaluations following two protocols: 1. Intra-class (P1): Training and testing are performed on the same motion classes, *e.g.*, jazziness. To ensure no overlap between training and testing, we split each dance class, allocating 80% of the images for training and the remaining 20% for testing. 2. Inter-class (P2): Images from the three types of dance classes are used for the training set, while the other five types are allocated to the test set. Due to page limits, more details about metrics and other setup are provided in the appendix.

Results and Analysis. Table 3 shows that HardMo-HMR performs much better than all existing models on HardMo-Foot P1, with an MPJPE-foot of **34.9mm**, which is 7× better than ProHMR [22] at 213.6mm, and almost 3× better than SOTA method, 4DHumans^b [10], at 88.1mm. What's more it achieves an MPJPE-body of 23.6mm, which is 5× better than the ProHMR, 1.7× better than 4DHumans^b. The above results confirm that our HardMo-Foot dataset not only effectively addresses the foot-hardcase issues but also benefits the body-part recovery. Similarly, HardMo-HMR also performs much better than all existing models on HardMo-Foot P2. These results demonstrate that after training on our dataset, the model doesn't merely learn the unique hardcase information of a particular class but indeed acquires a generalizable knowledge about foot movements. Moreover, we also perform training for HardMo-HMR without the optimized label, the comparison results demonstrating the necessity of optimized labels for foot recovery. More details are in appendix.

5.2.2 Benchmark on HardMo-Hand

Similar to the HardMo-Foot, we conduct comparison experiments with existing methods on HardMo-Hand. The detailed settings are attached in the appendix.

Results and Analysis. Table 4 shows that HardMo-HMR performs much better than all existing models on the HardMo-Hand dataset, with a PCK@0.01-hand of **0.36**. This is seven 7× better than ProHMR [22] at 0.03, and almost 3× better than SOTA method, 4DHumans^b [10], at 0.11. Moreover, it achieves a PCK@0.01-body of 0.54, which is almost 5× better than the ProHMR at 0.12, and is only lower than that of 4DHumans^b by a mere 0.02. The

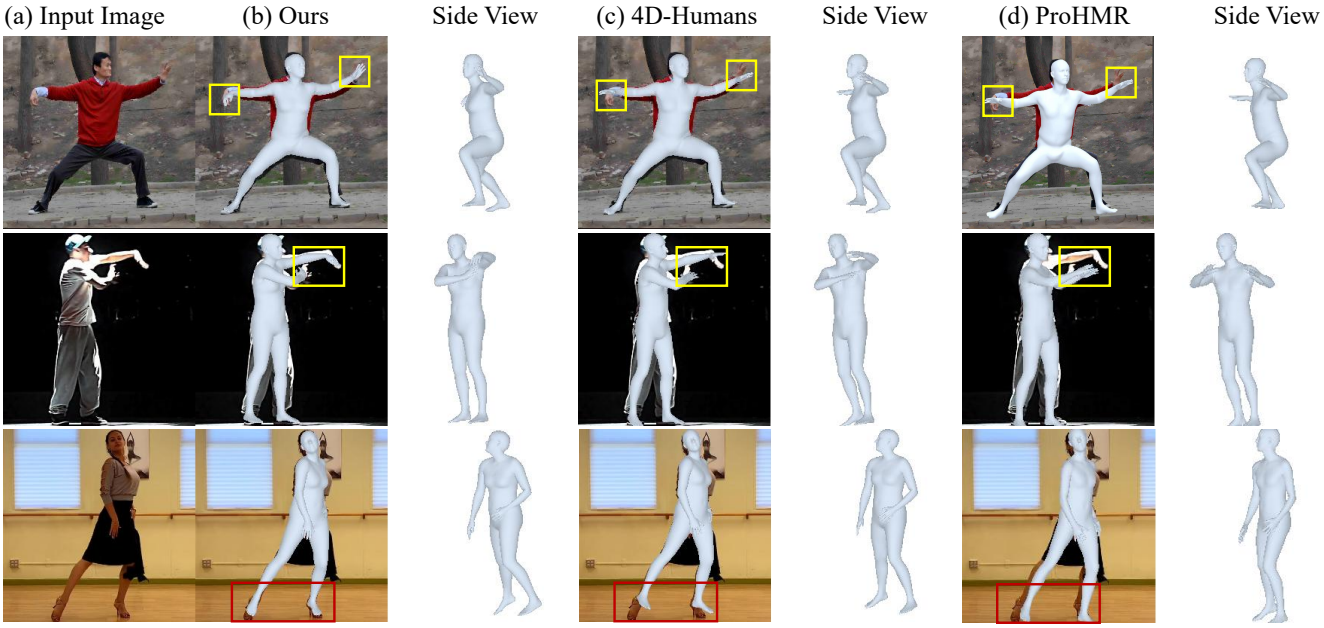

Figure 5. Qualitative Comparisons between HardMo-4DHumans and other leading methods. As shown in the figure, both ProHMR [22] and 4DHumans [10] fail to recover the correct hand-wrist posture. In contrast, after being finetuned on the hardcase subset of our HardMo dataset, HardMo-4DHumans can resolve these hardcases perfectly.

Method	PCK@0.01↑		PCK@0.05↑	
	Body	Hand	Body	Hand
ProHMR [22]	0.12	0.03	0.75	0.47
CLIFF [24]	0.32	0.08	0.93	0.64
HardMo-HMR	**0.54**	**0.36**	**0.98**	**0.98**
4DHumans^a [10]	0.15	0.04	0.93	0.49
4DHumans^b [10]	0.56	0.11	**0.99**	0.63
HardMo-4DHumans†	**0.61**	0.46	**0.99**	**0.99**
HardMo-4DHumans	0.58	**0.54**	**0.99**	**0.99**

Table 4. Reconstruction error on HardMo-hand. †: trained with mixture of HardMo-Foot and HardMo-Hand.

above results demonstrate that our HardMo-Hand dataset effectively addresses the hand-hardcase issues and benefits the body-part recovery.

Qualitative Results. In Fig. 5, we provide the qualitative results for tackling foot and hand hardcase problems. We compare HardMo-4DHumans with existing methods. The front view reveals that ProHMR [22] and 4DHumans [10] continue to struggle with severe hardcase issues. In contrast, HardMo-4DHumans has resolved these intrinsic hardcase problems perfectly.

6. Conclusion

Although existing methods perform well on benchmarks, they often struggle in real-world scenarios like dance and martial arts. Most of them suffer from severe misalignment in these cases where they tend to recover the posture of hands and feet at a T-pose. To bridge the gap between the research and application, we present HardMo, a large-scale dataset specially designed for complex motions. HardMo contains over 7 million images with 2D keypoints and SMPL [26] annotations. To further investigate the misalignment of feet and hands, we further filter two subsets, HardMo-Foot and HardMo-Hand dataset. Each subset is accompanied by optimized SMPL annotations. To efficiently build such datasets, we develop a efficient pipeline for automatic annotation, hardcase mining, and label optimization. Extensive experiments demonstrate the efficacy of the proposed pipeline and HardMo datasets. After training on HardMo, HMR [18], an early pioneering method, can even outperform the current SOTA, 4DHumans [10], on our benchmarks.

Limitation and Future Works. This paper only focuses on the misalignment of feet and hands. Apart from such hardcases, there still exist some other significant challenges, such as self-occlusion, that deserve further exploration. We leave this as our future work. From the data-driven perspective, we hope that our solution would enhance the mocap method in real-world scenarios and encourage more research centered around practical, real-world settings.

Acknowledgements. This work was supported by the National Key R&D Program of China (No. 2022ZD0116500), the National Natural Science Foundation of China (No. U21B2042, No. 62320106010), the InnoHK program, and in part by the 2035 Innovation Program of CAS.

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. 2, 3

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 3

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 3

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong 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. 2, 3

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*. Springer International Publishing, 2016. 3, 4, 5, 6

[6] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021. 3

[7] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pages 557–577. Springer, 2022. 3

[8] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1964–1973, 2021. 3

[9] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 2, 4

[10] 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, 2, 3, 4, 6, 7, 8

[11] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 6

[12] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1381–1388. IEEE, 2009. 3

[13] 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, 3

[14] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose, 2023. 2, 4

[15] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 2023. 4

[16] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 3

[17] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2020. 4, 5

[18] 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. 2, 3, 6, 8

[19] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 6

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020.

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

[22] 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. 2, 3, 6, 7, 8

[23] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 2, 3

[24] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 2, 3, 6, 7, 8

[25] 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. 2, 3

[26] 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. 2, 3, 4, 5, 6, 8

[27] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[28] 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. 2, 3

[29] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 2, 3

[30] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 3

[31] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1485–1495, 2022. 3

[32] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 3

[33] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023. 4

[34] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021. 3

[35] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4–27, 2010. 3

[36] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision*, pages 572–589. Springer, 2022. 3

[37] 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

[38] 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)*, 2018. 2, 3

[39] Jiong Wang, Fengyu Yang, Wenbo Gou, Bingliang Li, Danqi Yan, Ailing Zeng, Yijun Gao, Junle Wang, and Ruimao Zhang. Freeman: Towards benchmarking 3d human pose estimation in the wild. *arXiv preprint arXiv:2309.05073*, 2023. 3

[40] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 6

[41] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21222–21232, 2023. 3

[42] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 3

[43] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2, 3

[44] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[45] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 2248–2255, 2013. 3