MMVP: A Multimodal MoCap Dataset with Vision and Pressure Sensors

He Zhang[*1], Shenghao Ren[*3], Haolei Yuan[*1],
Jianhui Zhao[1], Fan Li[1], Shuangpeng Sun[2], Zhenghao Liang[4],
Tao Yu[†2], Qiu Shen[†3], Xun Cao[†3]

[1]Beihang University, [2]Tsinghua University, [3]Nanjing University, [4]Beijing Weilan Technology Co., Ltd.

Abstract

Foot contact is an important cue for human motion capture, understanding, and generation. Existing datasets tend to annotate dense foot contact using visual matching with thresholding or incorporating pressure signals. However, these approaches either suffer from low accuracy or are only designed for small-range and slow motion. There is still a lack of a vision-pressure multimodal dataset with large-range and fast human motion, as well as accurate and dense foot-contact annotation. To fill this gap, we propose a Multimodal MoCap Dataset with Vision and Pressure sensors, named MMVP. MMVP provides accurate and dense plantar pressure signals synchronized with RGBD observations, which is especially useful for both plausible shape estimation, robust pose fitting without foot drifting, and accurate global translation tracking. To validate the dataset, we propose an RGBD-P SMPL fitting method and also a monocular-video-based baseline framework, VP-MoCap, for human motion capture. Experiments demonstrate that our RGBD-P SMPL Fitting results significantly outperform pure visual motion capture. Moreover, VP-MoCap outperforms SOTA methods in foot-contact and global translation estimation accuracy. We believe the configuration of the dataset and the baseline frameworks will stimulate the research in this direction and also provide a good reference for MoCap applications in various domains. Project page: https://metaverse-ai-lab-thu.github.io/MMVP-Dataset/

1. Introduction

Human motion capture is an important foundation for motion analysis, behavior understanding, and pose generation, with a wide range of applications in AR/VR, disease diagnosis, robot manipulation, sports training, etc. In recent years, human motion capture based on computer vision technology has been developing rapidly, which has mainly
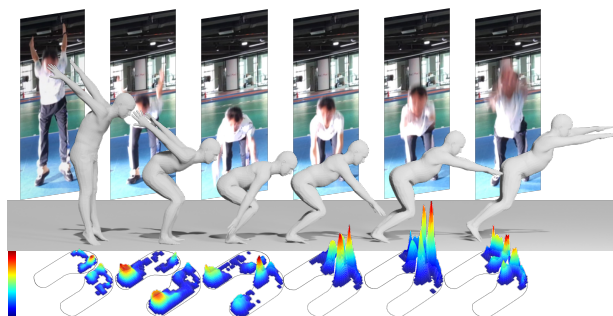

Figure 1. **MMVP** is a multimodal dataset that provides monocular RGBD video and accurate foot pressure (contact) of large-range and fast human motion.

benefited from the development of deep learning technology as well as, more importantly, the release of a large number of human motion capture datasets. Most of the existing datasets focus on human shape and pose annotations. Given pairs of image and shape and pose annotations for training, learning-based methods can infer plausible human motion from visual signals alone [5, 12, 23–26, 30, 37, 45, 56]. However, there are still many limitations when only using pose and shape annotations for supervision, resulting in artifacts such as foot drifting and erroneous global translation&rotation, etc.

To solve the above problem, recent datasets introduce additional annotations, mainly contacts, such as foot-contact, body-scene contact [15, 18] or self-contact [10, 34], to enhance additional physical constraints for visual motion capture. Among these contact annotations, the most classical one is the foot-contact annotation. With foot contact annotation, Rempe *et al.* [40] has inferred the 0-1 foot-contact label for the foot joints from an image, which successfully restricts the foot drifting artifacts and realizes much more accurate global translation results. It is worth mentioning that the labeled contact information can not only improve the accuracy of motion capture but also be used as important cues to directly improve the rationality of motion analysis [43, 44, 52] and even motion generation [16].

---
[*]Equal contribution.
[†]Corresponding author.

However, it is very difficult to accurately annotate the contact. Most of the existing datasets use purely visual matching and distance thresholding for contact annotation [14, 15, 18]. Specifically, these datasets usually require a pre-scanning of the 3D scene, and then, based on the image observations, use an optimization method to fit the parameterized human model (SMPL [32] or SMPL-X [38]) to the 3D scene, and finally the GT (ground truth) contact was annotated by thresholding the distance between the vertices of the SMPL model and the 3D scene. Since this is an in-direct annotation strategy, the accuracy is inevitably affected by the pose estimation error, shape estimation error, and even scene scanning error. The situation becomes even worse under fast and large-range motion when the contact time is short and the foot motion is sophisticated. In short, when the purely visual fitting error is larger, it is impractical to set a fixed distance threshold for accurate and detailed foot contact annotation. As a result, the foot contact annotation of these datasets is not precise enough and the annotation granularity is relatively coarse.

On the other end of the spectrum, in the field of health examination and clinical diagnosis, the use of specialized pressure sensors, such as the INSOLE, to obtain plantar pressure has been a proven and feasible option [11, 41, 50]. However, these datasets often lack the simultaneous acquisition of visual signals, making it difficult to apply to CV tasks. The most related multimodal datasets to ours are PSU-TMM100 [42] and MOYO [48]. Jesse *et al*. collected the PSU-TMM100 dataset with 2 RGB cameras and a pair of insole. However, this dataset is only designed for analyzing the stability of motion when humans perform Taiji Quan, so the movements are slow and the motion categories are relatively simple. Most recently, MOYO [48] uses multiple RGB sensors with a fixed position force plate for capturing yoga and analyzing the stability. However, the force plate restricts the pressure acquisition area for capturing large-range motion.

As a result, there is still a lack of a motion capture dataset that can simultaneously take into account a large range of motion, fast body movements, and visual-pressure synergistic acquisition, as well as accurate and dense foot-contact annotation. To fill this gap, we design a vision-pressure synchronized multimodal human motion capture system and acquire a new dataset named Multimodal MoCap dataset with Vision and Pressure Sensors (MMVP), as shown in Fig. 1. Specifically, we captured single-view RGBD sequences and plantar pressure under a large range of rapid movements with sophisticated footwork. Based on the high-end insole sensor, we can obtain precise and dense foot pressure and contact annotations which are significantly more accurate and detailed than previous datasets. To fully utilize the synchronized RGBD and dense foot-contact signals for motion capture, we further design a novel RGBD-

P SMPL fitting method based on multimodal input, which is far more accurate than vision-only methods, especially around foot regions. Finally, based on the new dataset and the fitted SMPLs, we propose and evaluate a baseline framework for monocular video-based human motion capture, VP-MoCap, which incorporates a per-vertex-level foot-pressure predictor (FPP-Net) and contact-based pose optimization strategy together to significantly improve the global translation and pose estimation accuracy.

In summary, our contributions are:

- We present MMVP, a novel multimodal human motion capture dataset that provides precise foot pressure and the most accurate dense contact annotations for large-range and rapid movements with detailed footwork.
- An RGBD-P SMPL fitting method is introduced to fully utilize the multimodal signals and achieve high-quality human motion capture results with plausible shape, stable foot movement, and accurate global translations. We also introduce a monocular RGB-based human motion capture baseline framework called VP-MoCap, which combines an FPP-Net and a joint optimization strategy.
- Comparisons with SOTA SMPL fitting and monocular MoCap methods demonstrate the effectiveness of our dataset, RGBD-P fitting strategy, and the VP-MoCap framework.

2. Related Work

2.1. Multimodal MoCap Datasets

The Human datasets [1, 3, 19, 33, 39, 55] have greatly promoted the development of pose detection. Nevertheless, these datasets only include visual signals and provide annotations for body shapes and poses. To further analyze human motion, multimodal datasets are needed. Tab. 1 summarizes existing multimodal datasets.

Some datasets introduce additional interaction annotations, such as human-scene contact [15, 18], self-contact [10, 34] or object contact [6], providing new insights for the study and analysis of human motion. These contacts can significantly reduce pose ambiguities in motion capture and enhance the understanding of visual images. Whether manually annotated or based on thresholds, these labels are derived solely from visual information, making it difficult to guarantee accuracy.

Some datasets incorporate signals from other modalities. LiDAR is typically applied in large scenarios, and its signal modality is similar to vision [8, 28, 51, 51]. Inertial Measurement Units (IMUs) can measure the accelerations and orientations of body parts. Thus the combination of vision and IMUs can solve heavy occlusion or extreme illumination [13, 49]. Pressure is a commonly used signal in human gait analysis, clinical rehabilitation assessment, and other biomedical fields. But medical datasets [11, 41, 50] often

Datasets	Visual Signal	Additional Signal	Additional Annotation
PROX [14]	S.V. RGBD	N/A	Scene-contact
RICH [18]	M.V. RGB	N/A	Scene-contact
HumanSC3D [10]	M.V. RGB	N/A	Self-contact
FlickrSC3D [10]	S.V. RGB	N/A	Self-contact
TUCH [34]	S.V. RGB	N/A	Self-contact
HOT [6]	S.V. RGB	N/A	Object contact
TotalCapture [49]	M.V. RGB	IMU	N/A
HPS [13]	M.V. RGB	IMU	N/A
LiDARHuman26M [28]	S.V. RGB	LiDAR	N/A
Schreiber et al. [41]	N/A	EMG, force plate	Foot pressure
Grouvel et al. [11]	N/A	IMU, insole, and force plate	Foot pressure
Zee et al. [50]	N/A	Insole	Foot pressure
MoYo [48]	M.V. RGBD	Force plate	Pressure and body contact
Ours	S.V. RGBD	Insole	Foot pressure and contact

Table 1. Comparison of existing multimodal human motion capture datasets. S.V.: single-view, M.V.: multi-view. Note that MoYo provides pressure and contact captured by a fixed-position force plate, so the capture range is limited. MMVP is the first dataset that provides accurate foot pressure and contact annotations with a large range and rapid body movements.

lack visual information. TMM100 [42] and MOYO [48] are the most representative vision & pressure datasets. The pressure can provide accurate contact information and physical features. However, both of these tend to favor slow and steady movements. Currently, there is still a lack of datasets that synchronize pressure and visual data to describe a wide range of fast body movements.

2.2. Human Pose Estimation

Existing methods have made remarkable progress in human pose and shape estimation [12, 21, 22, 29, 31, 60]. While single image and monocular pose estimation has received more widespread attention [4, 9, 23, 24, 26, 27, 30, 38, 59]. However, these methods solely focus on the correlation between image features and human pose, neglecting the connection between subjects and their environment. Consequently, this limitation often results in visually and physically implausible outcomes in the world frame, such as jitter, penetration, sliding, and so on.

LIDAR is employed to localize the global position of distant targets. LiDARCap [28], HSC4D [7], and Sloper4d [8] employ LIDAR data to estimate global human poses. CIMI4D [51] leverages LiDAR to acquire the human-scene contact information and regress the human pose for off-grounded motions. Due to the distance of the subjects, these methods often lack fine-grained motion details.

Some approaches have explored the fusion of RGB data with IMU data to enhance the accuracy and robustness of pose estimation. Matthew et al. [49] incorporates Long Short-Term Memory (LSTM) networks to fuse IMU tracking data and visual pose embeddings to improve the accuracy of pose estimation. HybridFusion [61] utilizes

IMU data to address the occlusion problems in a real-time monocular MoCap system. Pan et al. [35] fuses RGB and IMU data using a complementary filter algorithm within the mocap framework to address the challenge of reliability in vision information. EgoLocate [53] leverages MoCap priors from inertial inputs to enhance the accuracy and robustness of the mocap system.

Contact provides strong interaction priors for pose estimation. Thanks to the contact-based datasets, many methods [10, 15, 18] can estimate contact from a single image. The contact helps not only reduce ambiguity and achieve higher accuracy in pose estimation [10, 14, 34] but also provides dynamic constraints for rigid body dynamics-related methods [43, 44, 52]. In addition to determining contact, pressure can also provide force information. Zhang et al. [57] estimate human pose from three depth cameras and a pair of pressure-sensing shoes. Tripathi et al. [48] enhance the stability in pose estimation by encouraging plausible floor contact and overlapping the center of pressure and the SMPL's center of mass.

3. MMVP Dataset

3.1. Data Collection and Pre-processing

We collect a multimodal dataset containing synchronized RGBD video and foot pressure data. For visual data, we record the RGBD video using an Azure Kinect camera at a frame rate of 30Hz. For pressure data, we use Xsensor pressure insoles (HX 210-510). Each Xsensor pressure insole contains 242 independent pressure sensors that accurately capture the changes in pressure, with a frame rate of up to 150Hz. Due to the hardware limitation, we cannot
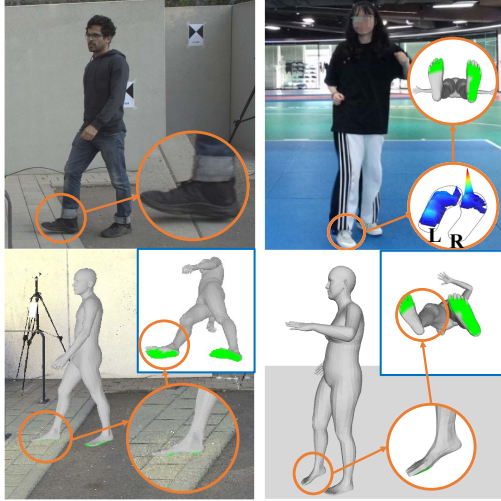
Figure 2. Comparison of the foot contact and pose annotations in RICH [18] (left) and MMVP (right). RICH fits the SMPL model for annotating the foot contact by distance thresholding, while MMVP incorporates dense pressure and contact directly for much more accurate SMPL fitting results. Note that due to the vision-only SMPL fitting error of RICH, the right foot, which is hanging in the air, was annotated as full contact with the ground.

synchronize the two streams of signals automatically, so we do manual synchronization before usage. The MMVP dataset covers up to 10 types of large-range and fast movements, including running, standing-long-jumping, and skipping, among other common types of exercises and dances. The majority of the subjects are teenagers, covering different age groups and both genders. For more details, please refer to the supplementary material.

3.2. Calculate Dense Foot Contact

Unlike defining contact on joints [40], dense contact [15] defines contact on the vertices of the human body model to describe detailed body-scene interaction. Existing methods [14, 18] usually first require scanning a 3D model of the scene and tracking the movements of subjects based on visual observations. Then annotate the dense contact by thresholding the distance between the body mesh and the scene. However, the inevitable pose estimation errors and scene reconstruction errors result in erroneous foot contact annotations, as shown in Fig. 2 (left).

Benefiting from the highly accurate pressure insole, we can calculate precise and dense foot contact annotations. To ease the usage of the dense pressure and contact information, we define it on the surface of the SMPL [32]. We select 192 vertices in total on the feet of the SMPL model and the dense contact label can be defined as $C \in \mathbb{R}^{192}$.

The foot pressure can vary dramatically depending on body weight and motion status. So it's hard to design a
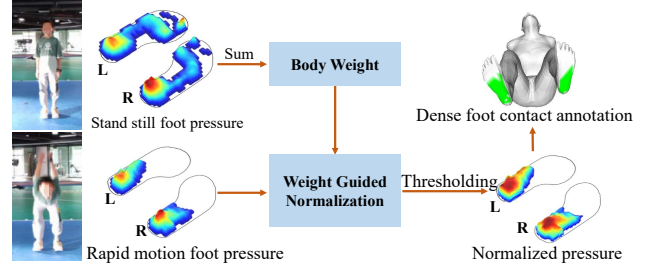

Figure 3. Illustration of the dense foot contact annotating method. From left to right are the reference image, original pressure, normalized pressure, and dense contact.

fixed-value threshold to annotate the dense contact. Luckily, with insoles, it's easy for us to get the accurate body weight of different subjects by simply summarizing the pressure at the beginning when the subject is standing still. Therefore we normalized and mapped the pressure to $[0, 1]$ as

$$P_{norm} = Sigmoid(P/w_s), \quad (1)$$

where P, P_{norm}, w_s are the originally captured pressure, normalized pressure, and body weight, respectively.

Finally, we empirically set the threshold to 0.5 to ensure a relatively strict contact annotation:

$$C = \begin{cases} 1, & P_{norm} \geq 0.5, \\ 0, & P_{norm} < 0.5. \end{cases} \quad (2)$$

Fig. 3 illustrates the calculation process of the dense foot contact annotation.

3.3. RGBD-P SMPL Fitting

As illustrated in Fig. 2, existing datasets calculate contact after fitting the body model to the pre-scanned scene with purely vision signals. Therefore, it is difficult to accurately maintain the interaction relationship between the human body and the scene during fitting. By constructing loss terms according to the ready-to-use contact information derived from the pressure signal, we can obtain more accurate pose and shape-fitting results. Specifically, we use SMPL [32] to numerically represent the body shape and pose. The SMPL model takes pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$, shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, global rotation $\boldsymbol{R} \in SO(3)$, and global translation $\boldsymbol{T} \in \mathbb{R}^3$ as input, producing a triangulated mesh of 6890 vertices. The k joints of the model are represented as $\boldsymbol{J}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{T}) \in \mathbb{R}^{k \times 3}$. To guarantee the shape-fitting performance for teenagers with various body scales in our dataset, we follow the operation in AGORA dataset [36], which introduces a blending parameter α to balance the shape difference between children and adults. According to AGORA, SMPL template $\boldsymbol{M}_A$ and SMIL [17] template $\boldsymbol{M}_C$ are blended as $M_{blend} =$
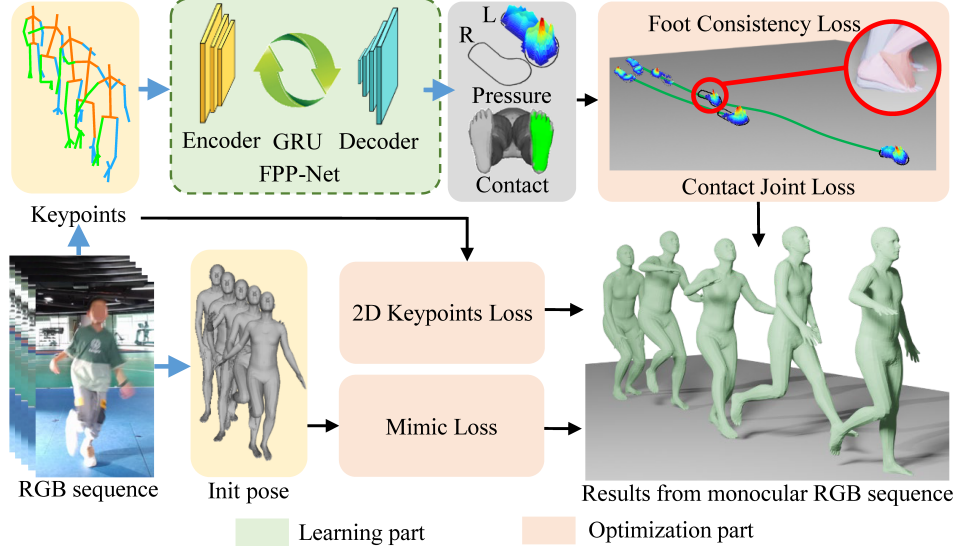
Figure 4. VP-MoCap pipeline. Given an RGB sequence, RTMPose [20] and CLIFF [30] are applied to detect 2D keypoints and regress the initial pose. FPP-Net predicts foot pressure distribution and dense foot contact with keypoint sequence. Guided by foot contact, joint optimization is applied to estimate pose and trajectory. (Green represents the learning part, while orange represents the optimization part.)

$\alpha \boldsymbol{M}_A + (1 - \alpha)\boldsymbol{M}_C$. The vertices of the blended model are denoted as $\boldsymbol{V}$.

We optimize Eq. (3) to fit the human model to the RGBD-P sequence with 3D depth fitting error E_{depth} introduced in Doublefusion [54] to align the depth point cloud and SMPL surface vertices, 2D projection fitting error E_{2d} and Gaussian Mixture Model(GMM) constraint E_{GMM} following SMPLify [4].

$$
\begin{aligned}
E(\boldsymbol{\theta}, \boldsymbol{T}, \alpha, \boldsymbol{\beta}) = & \lambda_{depth}E_{depth} + \lambda_{C\_dense}E_{C\_dense} + \\
& \lambda_{2d}E_{2d} + \lambda_{C\_temp}E_{C\_temp} + \\
& \lambda_{GMM}E_{GMM}.
\end{aligned} \tag{3}
$$

To fully utilize the annotated dense foot contact to eliminate foot-ground drifting, we introduce a dense contact loss E_{C_dense} in Eq. (4).

$$
E_{C\_dense} = \sum_{C=1} \|\Pi_{floor}(\boldsymbol{V}_{foot})\|, \tag{4}
$$

where $\Pi_{floor}(\boldsymbol{V}_{foot})$ denotes the distance between the ground and the foot surface vertex in contact.

Furthermore, to maintain the temporal smoothness of the foot fitting results during contact, we extend the human mesh to incorporate foot planes parallel to the pseudo ground. Each item in the plane corresponds to one vertex in the foot surface. These planes can be controlled in the same manner as human template vertices. Consequently, we design E_{C_temp} as Eq. (5) to enhance the temporal consistency of the contacted plane vertices. Please refer to the supplementary video for more details.

$$
E_{C\_temp} = \sum_{C_{t-1}=1 \cap C_t=1} \|(\boldsymbol{V}_{t-1} - \boldsymbol{V}_t)_{planes}\|_2, \tag{5}
$$

where $C_{t-1} = 1 \cap C_t = 1$ denotes the intersection between the sets of foot plane vertices in frame $t-1$ and frame t with a contact label of 1.

4. Method: VP-MoCap

Previous monocular RGB-based human mesh recovery methods [30] usually suffer from translation drifting, foot sliding, and foot-plane penetration, which make them unstable when handling relatively large-range and rapid movements. In the RGBD-P SMPL fitting section, we observed a much more accurate and plausible pose and translation by using depth and pressure signals for joint optimization. Thus, we further explore a more general question: given only monocular RGB sequences, can we also achieve stable and physically plausible human motion capture results for large-range and rapid movements with inferred dense contact and 3D scene?

Therefore, we first construct FPP-Net to estimate foot contact and then use Zoedepth [2] to estimate the 3D ground from a monocular RGB sequence. Finally, with foot contact and ground depth, we try to obtain better pose and translation through optimization using only the monocular RGB sequence. The overall framework is summarized in Fig. 4.

4.1. FPP-Net

To estimate foot pressure distribution and foot contact from an RGB sequence, we design a foot pressure predictor (FPP-Net), which is illustrated in the green dashed box in Fig. 4. We assume that foot pressure and contact are more related to body pose and motion dynamic state. Thus we follow Physcap [43] and Jesse *et al.* [42] to use sequential 2D keypoints for contact prediction. The goal of this strategy is to also decouple the network training process from the dataset-capturing environments (when compared with direct image regression methods like BSTRO [18] and DECO [47]) thus significantly enhancing the generalization capacity. The 2D keypoints sequence $\mathcal{J}^{2d} = \{j_1^{2d}, j_2^{2d}, ...j_t^{2d}\}$ is detected by RTMPose [20]. The task can be formulated as

$$\widetilde{C} = f(\mathcal{J}^{2d}). \tag{6}$$

We first use a encoder f_1 to extract motion feature $feat_t^{motion} \in \mathbb{R}^{2048}$ for each frame $f_1(j_1^{2d})$, $f_1(j_2^{2d})$, ..., $f_1(j_t^{2d})$. Then the motion feature is fed to a Gated Recurrent Unit (GRU) layer f_2 to yield temporal feature $feat_t^{temp} \in \mathbb{R}^{484}$. At last, a Multilayer Perceptron (MLP) is applied as the decoder to regress the final contact label $\widetilde{C} \in \mathbb{R}^{192}$.

Our training loss is

$$\mathcal{L} = \mathcal{L}_{cont} + \mathcal{L}_{press}. \tag{7}$$

Contact loss $\mathcal{L}_{cont}$ is the binary cross entropy loss between the ground truth contact and the predicted contact $\widetilde{C}$ [18]. Pressure loss $\mathcal{L}_{press}$ is the mean squared error loss between the ground truth pressure and the predicted pressure [42].

4.2. Pose and Translation Optimization

Since we pay more attention to human body pose and global translation, we do not optimize the intrinsic parameter K, shape $\tilde{\beta}$, and global rotation $\tilde{R}$; instead use weak-perspective projection to estimate K and adopt $\tilde{\beta}$, $\tilde{R}$, and initial pose $\tilde{\theta}$ from CLIFF [30]. It is important to note that an image sequence corresponds to a specific individual. Therefore, we fix an average of $\tilde{\beta}$ for the same person across the entire image sequence. To simplify the notation, we use $J(\theta, T)$ to represent $J(\theta, \tilde{\beta}, \tilde{R}, T)$.
2D keypoints loss. Due to large-range and rapid movements, our image sequence exhibits increased blur and occlusions, impacting the accuracy of the inferred pose parameter priors, $\tilde{\theta}$. To improve the accuracy of θ, we construct 2D keypoints loss, which serves to penalize the weighted distance between the reprojected SMPL joints and the corresponding estimated 2D keypoints $\mathcal{J}^{2d}$:

$$E_{2d} = \sum_i c_i \| \mathcal{J}_i^{2d} - \Pi_K J(\theta, T)_i \|_2^2, \tag{8}$$

where Π_K is the projection process from 3D to 2D through the intrinsic parameter K. The weight c is representative of the confidence associated with each estimated 2D keypoint.
Mimic loss. We also define a mimic loss to quantify the similarity of the pose and the initial pose,

$$E_p = \| \theta - \tilde{\theta} \|_2^2. \tag{9}$$

Contact joint loss. To recover the positional relationship between the human and the ground, we utilize the off-the-shelf depth estimation model ZoeDepth [2] to obtain a depth map for each RGB image and use the corresponding point cloud P to construct the 3D contact trace on the ground. Therefore, the spatial translation constraint term can be defined by penalizing the 3D distance between the SMPL foot joint $J(\theta, T)_{foot}$ and the contact point cloud $P(\mathcal{J}_{foot}^{2d})$:

$$E_{3d} = \sum \| P(\mathcal{J}_{foot}^{2d}) - J(\theta, T)_{foot} \|_2^2. \tag{10}$$

Foot consistency loss. Meanwhile, we assume that the foot joints' position of two consecutive frames is consistently close during contact. At the same time, the jitter of 2D keypoints $\mathcal{J}_{foot}^{2d}$ will cause the corresponding contact joint on the ground to shift significantly. To solve the above problems, we define a temporal constraint term as

$$E_t = \sum \| J(\theta^t, T^t)_{foot} - J(\theta^{t-1}, T^{t-1})_{foot} \|_2^2, \tag{11}$$

Combining all the above constraints Eqs. (8) to (11), the total objective function can be summarized as:

$$\arg \min_{\theta, T} E_{2d} + \lambda_p E_p + \lambda_{3d} E_{3d} + \lambda_t E_t, \tag{12}$$

where λ_p, λ_{3d} and λ_t are corresponding weights.

5. Experiments

5.1. Comparison of Ground Truth Registration

We compare our RGBD-P fitting method with two existing RGBD methods: PROX [14] and LEMO [58]. Both of these methods consider human-scene contact in pose fitting. We employ the 3D depth error E_{3d} to quantify the alignment between human models and depth observations. To evaluate the interaction between the model and the scene, we calculate the model contact by setting a unified threshold and compare the Mean Foot-Contact Error (MFCE), F1 score, and IOU with the ground truth contact. Tab. 2 demonstrates the effectiveness of the proposed method.

5.2. Comparison of Foot Contact Estimation

We apply precision, recall, F1 score, and IOU as evaluation metrics and conduct a quantitative comparison with BSTRO [18], BSTRO[FT], and DECO [47] on MMVP-test

Method	$E_{3d}\downarrow$	M.↓	F1↑	IOU↑
PROX [14]	0.045	9.348	0.475	0.439
LEMO [58][stage 2]	0.150	8.199	0.423	0.433
LEMO[stage 3]	0.175	8.231	0.377	0.399
Ours	0.043	7.968	0.518	0.507

Table 2. Evaluation of pose and shape registration on MMVP. M.: MFCE.

Method	prec.↑	recall↑	F1↑	IOU↑
BSTRO [18]	0.212	0.755	0.318	0.205
BSTRO[FT]	0.521	0.531	0.496	0.496
DECO [47]	0.112	0.853	0.191	0.112
Ours	0.529	0.586	0.532	0.522

Table 3. Quantitative comparison of contact estimation on MMVP-test.
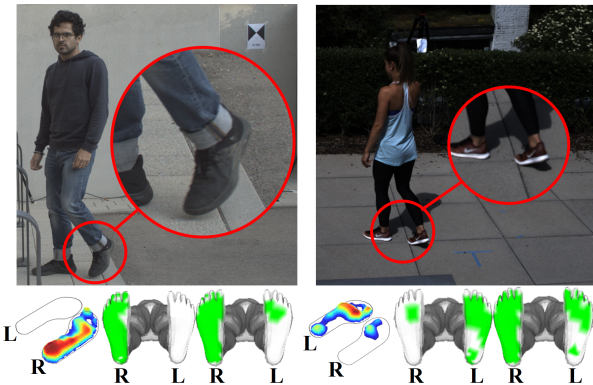

Figure 5. Qualitative comparison of contact estimation. From left to right, the second row includes: foot pressure distribution predicted by our method, foot contact predicted by our method, and contact predicted by BSTRO[FT].

Methods	M. ↓	PM. ↓	PVE ↓	PF. ↓	Traj ↓
CLIFF [30]	86.8	58.7	110.9	105.7	211.4
TRACE [46]	96.8	64.4	118.6	125.4	598.0
SMPLer-X [5]	92.2	51.1	112.9	121.8	679.7
Ours	83.0	56.0	110.6	91.9	129.3
(w/o L_t)	87.0	57.9	114.0	97.3	133.4
(w/o L_{3d}, L_t)	84.1	57.0	108.2	98.9	184.1

Table 4. Evaluation of pose and translation estimation on MMVP. M.: MPJPE, PM. : PMPJPE, PF. : PVE (Feet).

the proposed method not only achieves more accurate contact predictions but also estimates reasonable foot pressure distributions. For more details, please refer to the supplementary material.

5.3. Pose and Translation Optimization Results

To assess the accuracy of our 3D pose estimation, we employ various metrics: MPJPE (Mean Per Joint Position Error), PMPJPE (Procrustes-aligned Mean Per Joint Position Error), PVE (Per Vertex Error), and PVE (Feet) which specifically focuses on the error of SMPL mesh vertices corresponding to the soles of the feet. For evaluating translation estimation, we utilize the offset distance of the pelvis point relative to the starting position, denoted as Traj.

Quantitative results. We conduct a comparative analysis of our method against CLIFF [30], TRACE [46], and SMPLer-X [5] on MMVP. As depicted in Tab. 4, our method exhibits satisfactory performance across all metrics. Notably, our method excels in Traj, indicating a significant improvement in acquiring more accurate and robust translation by leveraging foot contact for SMPL fitting.

Qualitative results. We compare 3D translation between our method, CLIFF [30], and TRACE [46] on MMVP (Fig. 6). In the Side-Stepping case, both CLIFF and TRACE exhibit drift in the depth direction, whereas our method demonstrates a consistent and stable translation in the depth direction. Similarly, in the Standing-Long-Jump and Throwing-Medicine-Ball case, the 3D translation estimated by CLIFF and TRACE manifests drift in the depth direction during the preparation stage. In contrast, our method avoids any depth-direction translation. At the bottom, we present the average absolute acceleration of foot joints which means the level of jitter. Our method achieves the smallest value, indicating superior foot stability.

5.4. Ablation Studies

To investigate the impacts of L_{3d} and L_t, we conduct an ablation experiment on the MMVP dataset. As shown in Tab. 4, the omission of L_t results in a marginal increase in all error metrics. This observation indicates that the tempo-

(Tab. 3). BSTRO[FT] represents the BSTRO fine-tuned on the MMVP training dataset. From Tab. 3, it is evident that our method achieves the best performance. It is worth mentioning that BSTRO and DECO, due to contact annotations in their training dataset (illustrated in Fig. 2), tend to predict feet as in contact for the majority of cases. As a result, their recall values are very high, while other metrics are undesirable. It is challenging to perceive the foot contact solely based on a single-frame image. By incorporating temporal motion information, we believe that's why the proposed method surpasses the fine-tuned BSTRO.

To validate the generalization, we also compare our method with BSTRO[FT] on public datasets. Due to the inaccurate contact labels in other datasets, we only conduct qualitative comparisons (Fig. 5). It can be observed that
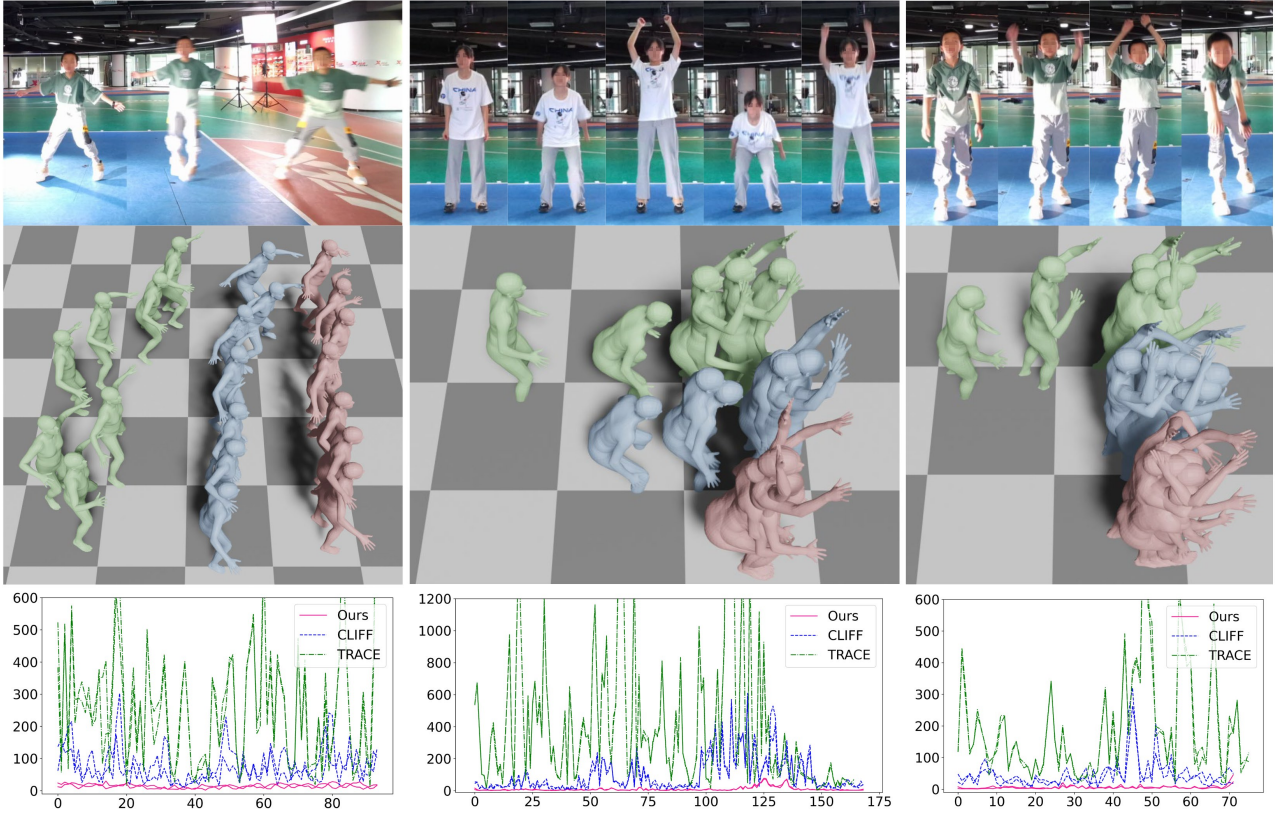
Figure 6. Qualitative comparison of 3D translation estimation in Side-Stepping (Left), Standing-Long-Jump (Middle) and Throwing-Medicine-Ball (Right) case. We estimate 3D translation stability and compare our results (Pink) with CLIFF [30] (Blue) and TRACE [46] (Green). The bottom row represents the acceleration of foot joints over time, which is used to measure the jitter of foot joints.

ral foot contact signals serve not only to improve the accuracy of translation but also to improve the accuracy of the pose. Without L_{3d}, MPJPE, PA-MPJPE, and PVE dropped slightly, but PVE (Feet) and Traj increased a lot. This is because when the 3D constraints of the feet and the ground are introduced, it will affect the 2D constraints to a certain extent and cause the accuracy of the pose to drop a little, but in exchange for the accuracy of the soles of the feet vertices and global translation. To attain enhanced translation, it is necessary to make a trade-off by sacrificing a certain degree of pose accuracy. Additionally, getting a better translation requires striking a balance between 2D and 3D constraints.

6. Conclusion

The interaction between the human body and the scene has received increasing attention in recent years. However, most existing works struggle to provide precise contact annotations. To tackle this problem, we collect MMVP, a novel vision-pressure dataset with a wide range of rapid movements. In addition, we present VP-MoCap, a monocular MoCap baseline framework that incorporates the prediction of dense pressure and contact for pose optimization.

Compared to solely relying on contact annotations, pressure signals contain richer dynamics information, particularly for fast and large-scale movements. We believe that our dataset holds great potential for future research.

Limitations and future work: Our method is mainly designed for static viewpoints, so it remains challenging to handle moving camera configurations. The existing MMVP has relatively homogeneous scenes, making it difficult for current visual networks to achieve scene generalization on this dataset. In the future, we plan to collect data from various outdoor scenes to expand the dataset. Additionally, it would be an intriguing research endeavor to learn human body dynamics solely from pressure sequences.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, pages 8387–8397, 2018. 2

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 5, 6

[3] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. Faust: Dataset and evaluation for 3d mesh registration. In *CVPR*, pages 3794–3801, 2014. 2

[4] 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 *ECCV*, pages 561–578, 2016. 3, 5

[5] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qing-ping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *NIPS*, 36, 2024. 1, 7

[6] Yixin Chen, Sai Kumar Dwivedi, Michael J Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *CVPR*, pages 17100–17110, 2023. 2, 3

[7] Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *CVPR*, pages 6792–6802, 2022. 3

[8] Yudi Dai, YiTai Lin, XiPing Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *CVPR*, pages 682–692, 2023. 2, 3

[9] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *3DV*, pages 792–804. IEEE, 2021. 3

[10] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *AAAI*, pages 1343–1351, 2021. 1, 2, 3

[11] Gautier Grouvel, Lena Carcreff, Florent Moissenet, and Stéphane Armand. A dataset of asymptomatic human gait and movements obtained from markers, imus, insoles and force plates. *Scientific Data*, 10(1):180, 2023. 2, 3

[12] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, pages 10884–10894, 2019. 1, 3

[13] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *CVPR*, pages 4318–4329, 2021. 2, 3

[14] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambigui-ties with 3d scene constraints. In *ICCV*, pages 2282–2292, 2019. 2, 3, 4, 6, 7

[15] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *CVPR*, pages 14708–14718, 2021. 1, 2, 3, 4

[16] Alexander Herzog, Stefan Schaal, and Ludovic Righetti. Structured contact force optimization for kino-dynamic motion generation. In *IROS*, pages 2703–2710. IEEE, 2016. 1

[17] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. Learning an infant body model from rgb-d data for accurate full body motion analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 792–800. Springer, 2018. 4

[18] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, pages 13274–13285, 2022. 1, 2, 3, 4, 6, 7

[19] 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 TPAMI*, 36(7):1325–1339, 2014. 2

[20] 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. *arXiv preprint arXiv:2303.07399*, 2023. 5, 6

[21] 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 *ICCV*, pages 3334–3342, 2015. 3

[22] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, pages 8320–8329, 2018. 3

[23] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 1, 3

[24] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5252–5262, 2020. 3

[25] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137, 2021.

[26] 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*, pages 2252–2261, 2019. 1, 3

[27] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, pages 3383–3393, 2021. 3

[28] Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture

with lidar point clouds. In *CVPR*, pages 20502–20512, 2022. 2, 3

[29] Kun Li, Qionghai Dai, and Wenli Xu. Markerless shape and motion capture from multiview video sequences. *IEEE TCSVT*, 21(3):320–334, 2011. 3

[30] 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 *ECCV*, pages 590–606. Springer, 2022. 1, 3, 5, 6, 7, 8

[31] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE TPAMI*, 35(11):2720–2735, 2013. 3

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 2, 4

[33] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, pages 5441–5450, 2019. 2

[34] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *CVPR*, pages 9990–9999, 2021. 1, 2, 3

[35] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. Fusing monocular images and sparse imu signals for real-time human motion capture. In *SIGGRAPH Asia*, pages 1–11, 2023. 3

[36] 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 *CVPR*, pages 13468–13478, 2021. 4

[37] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018. 1

[38] 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 *CVPR*, pages 10975–10985, 2019. 2, 3

[39] 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 *CVPR*, pages 9054–9063, 2021. 2

[40] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *ECCV*, pages 71–87. Springer, 2020. 1, 4

[41] Céline Schreiber and Florent Moissenet. A multimodal dataset of human gait at different walking speeds established on injury-free adult participants. *Scientific data*, 6(1):111, 2019. 2, 3

[42] Jesse Scott, Bharadwaj Ravichandran, Christopher Funk, Robert T Collins, and Yanxi Liu. From image to stability: Learning dynamics from human pose. In *ECCV*, pages 536–554. Springer, 2020. 2, 3, 6

[43] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM TOG*, 39(6):1–16, 2020. 1, 3, 6

[44] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM TOG*, 40(4): 1–15, 2021. 1, 3

[45] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 1

[46] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *CVPR*, pages 8856–8866, 2023. 7, 8

[47] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *ICCV*, pages 8001–8013, 2023. 6, 7

[48] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *CVPR*, pages 4713–4725, 2023. 2, 3

[49] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *BMVC*, pages 1–13, 2017. 2, 3

[50] Tim J van der Zee, Emily M Mundinger, and Arthur D Kuo. A biomechanics dataset of healthy human walking at various speeds, step lengths and step widths. *Scientific data*, 9(1): 704, 2022. 2, 3

[51] Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. In *CVPR*, pages 12977–12988, 2023. 2, 3

[52] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *CVPR*, pages 13167–13178, 2022. 1, 3

[53] Xinyu Yi, Yuxiao Zhou Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM TOG*, 42(4), 2023. 3

[54] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*, pages 7287–7296, 2018. 5

[55] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, pages 5746–5756, 2021. 2

[56] 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 *ICCV*, pages 11446–11456, 2021. 1

[57] Peizhao Zhang, Kristin Siu, Jianjie Zhang, C Karen Liu, and Jinxiang Chai. Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture. *ACM TOG*, 33(6):1–14, 2014. 3

[58] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *CVPR*, pages 11343–11353, 2021. 6, 7

[59] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, pages 7376–7385, 2020. 3

[60] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Lightweight multi-person total motion capture using sparse multi-view cameras. In *ICCV*, pages 5560–5569, 2021. 3

[61] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *ECCV*, pages 384–400, 2018. 3