Improving 3D Foot Motion Reconstruction in Markerless Monocular Human Motion Capture

Tom Wehrbein Bodo Rosenhahn
L3S - Leibniz University Hannover, Germany
wehrbein@tnt.uni-hannover.de

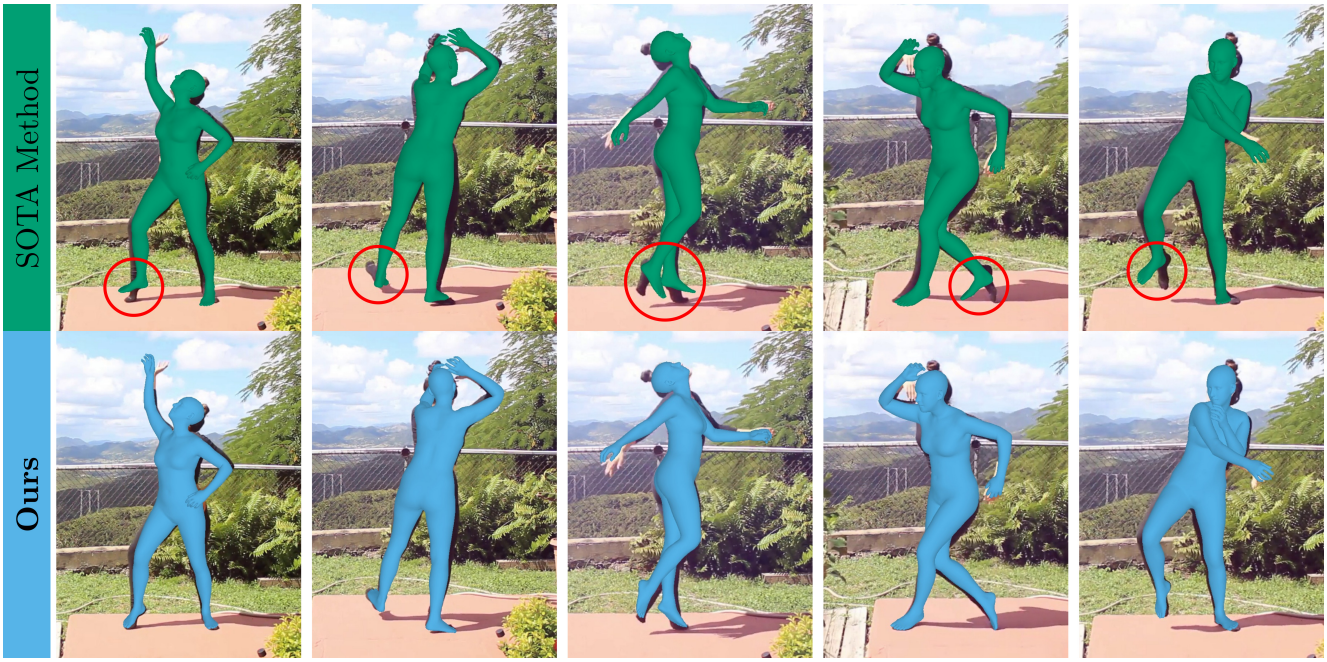

Figure 1. State-of-the-art 3D human motion recovery methods like GVHMR [42] *(top)* fail to capture complex 3D foot movement when given in-the-wild videos. We identify this to be mainly an issue of inaccurate and insufficient video training data. To address this, we introduce FootMR, a Foot Motion Refinement method that leverages large-scale motion capture data to learn lifting 2D foot keypoint sequences to 3D. By effectively resolving ambiguities in 2D-to-3D mapping, FootMR *(bottom)*, when combined with an existing 3D human recovery model, generates realistic and accurate 3D foot motion, significantly outperforming previous work.

Abstract

State-of-the-art methods can recover accurate overall 3D human body motion from in-the-wild videos. However, they often fail to capture fine-grained articulations, especially in the feet, which are critical for applications such as gait analysis and animation. This limitation results from training datasets with inaccurate foot annotations and limited foot motion diversity. We address this gap with FootMR, a Foot Motion Refinement method that refines foot motion estimated by an existing human recovery model through lifting 2D foot keypoint sequences to 3D. By avoiding direct image input, FootMR circumvents inaccurate image–3D annotation pairs and can instead leverage large-scale motion capture data. To resolve ambiguities of 2D-to-3D lifting, FootMR incorporates knee and foot motion as context and predicts only residual foot motion. Generalization to extreme foot poses is further improved by representing joints in global rather than parent-relative rotations and applying extensive data augmentation. To support evaluation of foot motion reconstruction, we introduce MOOF, a 2D dataset of complex foot movements. Experiments on MOOF, MOYO, and RICH show that FootMR outperforms state-of-the-art methods, reducing ankle joint angle error on MOYO by up to 30% over the best video-based approach. Our code and dataset are available for research purposes at twehrbein.github.io/footmr-website/.

1. Introduction

Accurately reconstructing the 3D motion of a person from monocular video has been a major research problem for several decades. The field has advanced substantially in recent years, and current methods can robustly recover promising human motion from casual videos. However, while the coarse body movement is usually reconstructed with high accuracy, they fail at reconstructing the fine-grained movement of the feet (see Fig. 1). We argue that accurately reconstructing nuanced foot motion is particularly important for many applications in sports, medicine, AR/VR, and animation. For example, when capturing the performance of a dancer, the movement of the feet plays a critical role in producing detailed and lifelike animation.

Our insight is that previous approaches fail because they rely on in-the-wild training data with inaccurate 3D foot annotations. Such pseudo-ground truth (pseudo-GT) annotations are generated by fitting a parametric body model to sparse 2D keypoints [8, 17, 25, 27, 34], sometimes combined with IMU data [52]. Although these fitting targets can constrain the coarse 3D body pose, details of the feet are often lost because keypoints are typically defined only for major body joints, extending no further than the ankle. With only a single keypoint for the ankle, the 3D pose of the foot is not sufficiently constrained, leading to inaccurate pseudo-GT fits as shown in Fig. 2. However, training on large-scale, diverse data is crucial for models to generalize well to different motions and scenes. Another challenge is that existing 3D human video datasets [12, 13, 33, 52] mainly contain subjects performing everyday activities with very little foot motion or synthetic humans without shoes [2, 51]. This limits the ability of models to generalize to complex foot movements typical of dance, ballet, and sports.

We address these challenges with *FootMR*, a *Foot Motion Refinement* method that refines foot motion estimated by an existing 3D human motion recovery model. Instead of directly using images as input, FootMR processes 2D foot keypoints from an off-the-shelf detector [20]. We use four keypoints per foot: big toe, small toe, heel, and ankle. During training, we synthetically generate 2D keypoints and ground truth 3D motion sequence pairs using both large-scale motion capture and video datasets. By avoiding direct image input, FootMR completely bypasses the dependency on inaccurate image-3D foot annotation pairs. Foot articulation in parametric 3D human body models is primarily defined by rotations of the ankle joint. Therefore, we train a model to learn lifting 2D foot keypoint sequences to 3D ankle rotations. Because 2D-to-3D lifting is inherently ambiguous and degenerates when input keypoints are noisy, we incorporate knee and initial ankle rotations estimated by an existing 3D human recovery model as additional input and predict only residual ankle rotations. The knee is the parent joint of the ankle along the kinematic chain and thus

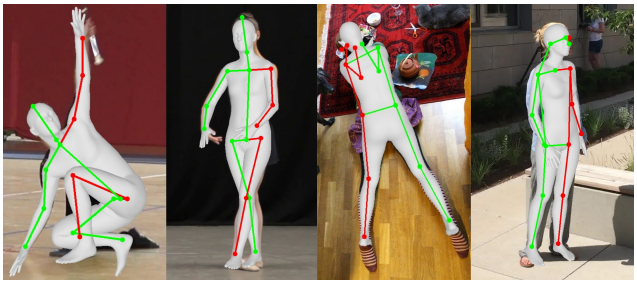

Figure 2. **Erroneous 3D foot annotations** in pseudo-GT fits generated by fitting 3D models to sparse keypoints. Images are from MPII [1], COCO [28], and 3DPW [52]. Please zoom in for details.

provides information about the space of feasible ankle rotations. Together with leveraging initial ankle estimates and the temporal motion context, this helps resolving ambiguity of the 2D-to-3D mapping. To further improve generalization to extreme foot poses, FootMR processes global rather than parent-relative joint rotations. Intuitively, because foot motion diversity in training datasets is limited, the range of possible 3D ankle rotations observed during training is very narrow when using parent-relative rotations. Additionally, we use heavy data augmentation by applying a random 3D rotation to the root orientation of all 3D poses of a sequence. This is possible because no images are used for training FootMR, and the 2D keypoints can be efficiently synthesized through projection with a virtual camera. Without using additional training data or having to meticulously fix pseudo-GT annotations, FootMR accurately captures intricate 3D motion of the feet (see Fig. 1). The refined ankle predictions are fused with the remaining body parameters estimated by a 3D human motion recovery model.

To support evaluation of foot motion reconstruction, we collect a new dataset with complex **MO**vements **O**f the **F**eet (MOOF). We record videos of individuals performing simple body movements with complex foot movements, *e.g.*, a person sitting on a chair doing ankle circles, and extend these recordings with dance and ballet videos collected online. In total, MOOF consists of 41 videos with annotated 2D foot keypoints. Experiments on MOOF and on the 3D datasets MOYO and RICH demonstrate that FootMR achieves significantly more accurate foot motion reconstruction than all competitors.

Our main contributions are summarized as follows:

- We propose FootMR, a method that leverages 2D foot keypoints to refine 3D foot motion estimated by a human recovery model.
- We show that 2D-to-3D foot motion lifting works robustly when using knee and initial foot motion as context.
- We collect MOOF, a new video dataset containing complex foot movements with annotated 2D foot keypoints.
- FootMR generalizes to extreme foot poses and outperforms all competitors on MOYO, RICH, and MOOF.

2. Related Work

2.1. Monocular Human Mesh Recovery

Reconstructing 3D human pose and shape from monocular images is most widely formulated as estimating the low-dimensional parameters of a statistical body model [16, 29, 35, 37, 61]. Pioneering work [3, 9, 10, 46] investigates optimization-based approaches by fitting the body parameters to image observations. Starting with HMR [18], direct regression methods based on deep learning became the leading paradigm. Many methods follow HMR in using a backbone to extract image features followed by a multilayer perceptron (MLP) that regresses body parameters [11, 27, 36, 59, 70]. Improved backbones [6, 8, 48] and better camera modeling [21, 23, 27, 36, 54, 55] played an important role in advancing reconstruction accuracy in recent years. As for all data-driven approaches, another key factor for accuracy and robustness is the availability of high-quality, large-scale datasets. Since such 3D human datasets, especially in the wild, are difficult to obtain, previous work focuses on generating 3D pseudo-GT annotations. This is done by fitting body parameters to 2D keypoints [8, 25, 36, 58] using SMPLify [3], or by fine-tuning a pretrained body regressor on the target images using 2D keypoints as weak supervision [17, 27, 34]. Although training with 3D pseudo-GT annotations is crucial for models to generalize well, we notice that they are often inaccurate for the feet (see Fig. 2) and thus lead to models with good body but poor foot pose reconstructions. CameraHMR [36] improves upon this by using a newly introduced dense surface keypoint detector to estimate more keypoints for pseudo-GT fitting. However, it fails to generalize to extreme foot poses not seen during training. Additionally, image-based models applied to frames of a video sequence often produce temporally inconsistent body poses and shapes.

Video-based approaches encode temporal information by jointly processing static features extracted from each frame. Earlier methods use convolutional [19] or recurrent encoders [5, 22, 31], while transformer architectures are employed by more recent methods [41, 53, 60]. Even more recently, several approaches [24, 26, 42, 45, 49, 57, 65, 68, 69] aim to recover global human motion to handle arbitrary moving cameras. While producing temporally coherent results and accurate body estimates, all methods fail to capture complex 3D foot movements because, similar to image-based methods, they rely on training data with inaccurate foot annotations. Additionally, existing 3D human video datasets are much more limited in diversity than image datasets, typically containing very little foot motion.

2.2. Foot Pose Estimation

Most existing human pose datasets provide only minimal foot annotations, limited to the position of the ankles. The first 2D foot keypoint dataset was released by Cao *et al*. [4] which extends a subset of COCO [28] by labeling three keypoints per foot: big toe, small toe, and heel. Similarly, Jin *et al*. [15] introduced COCO-WholeBody for 2D human whole-body pose estimation. Enabled by these datasets, recent 2D detectors [14, 20, 62–64] can accurately and robustly detect foot keypoints in addition to body keypoints.

To reconstruct foot keypoints in 3D, previous work [16, 72, 73] relies on multi-view optimization using corresponding 2D detections. Zhu *et al*. [72] provide their reconstructions for two existing multi-view datasets [13, 16]. Zhuo *et al*. [73] construct annotations for training a human mesh recovery model, but do not release the annotations or the model. Relying on multi-view indoor data is very restricting and leads to a lack of diversity in scenes and actors, limiting robustness of data-driven methods. In contrast, by refining foot motion decoupled from the rest of the body, FootMR only requires 2D foot keypoints and can thus leverage large-scale motion capture data together with in-the-wild video datasets without suffering from inaccurate foot annotations. During inference, FootMR benefits from the accuracy and robustness of recent 2D whole-body keypoint detectors.

Several recent works [38, 39, 43, 66, 67, 74] focus on the physical plausibility of 3D human motion to mitigate artifacts such as foot sliding or foot-floor penetration. However, they do not address the accuracy or evaluation of foot motion reconstruction, which is the focus of this work.

3. Method

Given a monocular video $\{I^{(l)}\}_{l=1}^{L}$ of length L, our objective is to recover 3D human motion that is accurate not only for the body but also for the feet. We adopt SMPL-X [37] to represent the 3D human body, which consists of J joint rotations $\{\theta^{(l)} \in \mathbb{R}^{J\times 6}\}_{l=1}^{L}$, shape parameters $\beta \in \mathbb{R}^{10}$, and translation $\{\tau_c^{(l)} \in \mathbb{R}^3\}_{l=1}^{L}$ in the camera space. We use the continuous 6D rotation representation proposed by [71] to represent 3D joint rotations. To parameterize foot articulation, the SMPL-X kinematic tree contains a joint for the ankle and forefoot. Since the ankle joint has the most influence on foot motion, and previous work fails to accurately capture it, we focus on improving motion reconstruction of the left and right ankles $\{\theta_{\text{ankle}}^{(l)} \in \mathbb{R}^{2\times 6}\}_{l=1}^{L}$.

To address the challenge of insufficient video data with accurate 3D foot annotations, we decouple the reconstruction process. First, a SMPL-X motion estimator is used to estimate 3D human motion including potentially erroneous initial ankle rotations. The initial ankle predictions are then refined by our proposed Foot Motion Refinement (FootMR) method, which utilizes 2D foot keypoints instead of raw images and thus avoids relying on inaccurate image-3D foot annotation pairs during training. An overview of our framework is shown in Fig. 3.

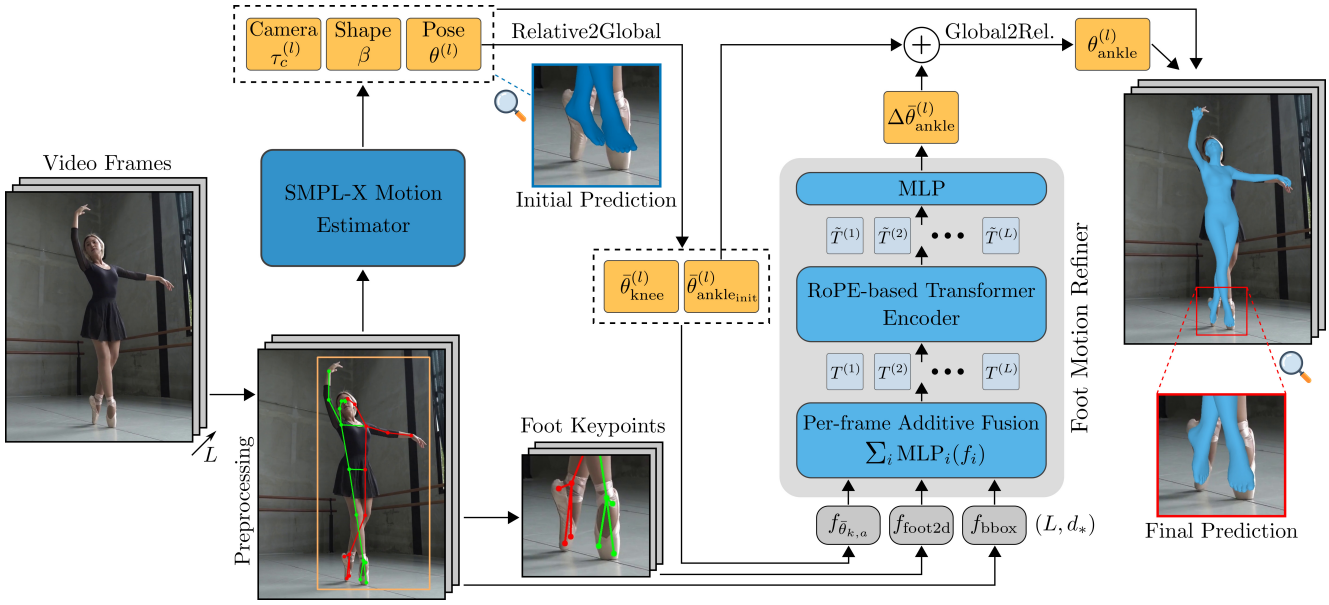

Figure 3. **Overview of our framework.** Given a monocular video, a SMPL-X motion estimator is employed to estimate 3D human motion including erroneous initial ankle rotations. Knee and ankle predictions are then transformed from parent-relative to global rotations and used together with 2D foot keypoints and bounding boxes as input for Foot Motion Refinement (FootMR). FootMR refines the initial ankle predictions by estimating residual rotations. After fusing the refined ankle rotations with the remaining body parameters, the final output of our framework is accurate and temporally coherent 3D human body *and* foot motion.

3.1. Foot Motion Refinement

Input and preprocessing. We design a transformer-based model to jointly refine initial left and right ankle rotations $\bar{\theta}^{(l)}_{\text{ankle}_{\text{init}}} \in \mathbb{R}^{2\times 6}$ by predicting residual rotations $\Delta\bar{\theta}^{(l)}_{\text{ankle}} \in \mathbb{R}^{2\times 6}$ from three types of input. The first is the 2D foot keypoint sequence $f_{\text{foot2d}} \in \mathbb{R}^{L\times 16}$ that consists of four keypoints per foot: big toe, small toe, heel, and ankle. We normalize the keypoints using the person's bounding box and provide the center and scale of the bounding boxes as second input $f_{\text{bbox}} \in \mathbb{R}^{L\times 3}$ [27]. This provides information about the location of the person in the original image which is important due to perspective effects. Keypoints that are not visible based on their confidence score are set to zero. The third input of FootMR is the global rotations of the left and right knee and ankle $f_{\bar{\theta}_{k,a}} \in \mathbb{R}^{L\times 24}$ predicted by a SMPL-X motion estimator. In SMPL-X, each joint rotation is defined relative to its parent joint in the kinematic tree. Let $\theta_i^{(l)} \in \mathbb{R}^6$ denote the parent-relative rotation of joint i in 6D representation. Following Zhou *et al.* [71], the corresponding rotation matrix is constructed by

$$\mathbf{R}_i^{(l)} = \text{rotmat}(\theta_i^{(l)}), \quad \mathbf{R}_i^{(l)} \in \text{SO}(3), \tag{1}$$

with the inverse mapping

$$\theta_i^{(l)} = \text{rot6d}(\mathbf{R}_i^{(l)}), \quad \theta_i^{(l)} \in \mathbb{R}^6. \tag{2}$$

Since relative knee rotations provide no useful information for ankle pose refinement, we transform them into global rotations in camera space. For a joint i, this is done by multiplying the rotations of all its ancestor joints $\mathcal{A}(i)$ along the kinematic chain, from the root to the target joint:

$$\bar{\mathbf{R}}_i^{(l)} = \Big(\overrightarrow{\prod_{j\in\mathcal{A}(i)}} \mathbf{R}_j^{(l)} \Big) \mathbf{R}_i^{(l)}. \tag{3}$$

The operator $\overrightarrow{\prod}$ denotes the ordered matrix multiplication along the chain, and $\bar{\theta}_i^{(l)} = \text{rot6d}(\bar{\mathbf{R}}_i^{(l)})$ is the resulting global rotation of joint i in 6D representation. Conditioning on global knee rotations is crucial, as they constrain feasible ankle rotations and help disambiguate lifting 2D keypoint sequences to 3D rotations. Initial ankle predictions act as a strong prior, and are especially important when 2D keypoints are noisy or missing. Using global rather than knee-relative ankle rotations leads to FootMR better generalizing to extreme ankle motions.

Network design. Our network design is heavily inspired by GVHMR [42]. The three per-frame input features $f^{(l)}_{\bar{\theta}_{k,a}}, f^{(l)}_{\text{foot2d}}, f^{(l)}_{\text{bbox}}$ are first independently mapped to the same dimension using dedicated MLPs. They are then combined through element-wise summation to create one unified token $T^{(l)} \in \mathbb{R}^{d_{\text{H}}}$ per frame. Subsequently, the resulting sequence is processed by multiple transformer encoder layers leveraging Rotary Position Embedding (RoPE) [47], which encodes relative positional dependencies instead of absolute positions. Relative positional embeddings together

with a further introduced attention mask enable FootMR to process sequences of arbitrary length in a single forward pass, without relying on autoregressive inference strategies such as sliding-window. The attention mask is constructed such that each token only attends to tokens within a W-frame neighborhood. Due to the design decision to have only one token per frame, FootMR is computationally efficient even for larger W and thus can effectively capture long-range dependencies. The output tokens $\{\tilde{T}^{(l)} \in \mathbb{R}^{d_\mathrm{H}}\}_{l=1}^{L}$ of the final transformer encoder layer are independently processed by an MLP to predict the residual global rotation of the left and right ankle $\Delta\bar{\theta}^{(l)}_{\text{ankle}} \in \mathbb{R}^{2\times 6}$. Due to its efficient design and because only a few keypoints and rotations are processed per frame instead of heavy image features, our foot motion refinement is lightweight and introduces only minor computational overhead.

Network output. The output of FootMR is simply added element-wise to the initial ankle predictions, producing the final global ankle rotations

$$\bar{\theta}^{(l)}_{\text{ankle}} = \bar{\theta}^{(l)}_{\text{ankle}_{\text{init}}} + \Delta\bar{\theta}^{(l)}_{\text{ankle}}. \tag{4}$$

To then convert the global to parent-relative ankle rotations, they must be multiplied by the inverse of the parent joint's rotation matrix:

$$\mathbf{R}^{(l)}_{\text{ankle}} = \left(\bar{\mathbf{R}}^{(l)}_{\text{knee}}\right)^{-1} \bar{\mathbf{R}}^{(l)}_{\text{ankle}}. \tag{5}$$

3.2. Training

We jointly train FootMR with the SMPL-X motion estimator from scratch. We find that this produces slightly better results than training FootMR on top of a pretrained fixed-weight SMPL-X regressor. We follow previous work [42, 44, 45] and synthesize input 2D keypoints during training by extracting 3D keypoints from the ground truth SMPL-X sequence, adding noise and projecting them onto a virtual camera. Since image features are not used as input, additional data augmentation for training FootMR can be applied. Specifically, for every sequence, we uniformly sample a random 3D rotation $\mathcal{R} \sim \mathcal{U}(\mathrm{SO}(3))$ and apply it to the ground truth and predicted root joint orientations. This results in augmented foot keypoints, bounding boxes, and predicted global knee and ankle rotations. By effectively simulating 3D foot motion in all possible orientations in 3D space, the variety of keypoints and rotations FootMR processes during training is heavily expanded. This helps the model to generalize well to diverse foot motions. Note that the random rotation data augmentation is only applied to the inputs of FootMR. The output does not need to be adjusted, as the updated global ankle rotations are always converted to parent-relative rotations that are independent of the orientation of the root.

We do not use any additional losses for training FootMR, but instead simply integrate its output into the losses of the SMPL-X motion estimation method. Typical motion reconstruction losses influenced by the ankle rotations are

$$\mathcal{L}_{\text{j3d}} = ||\mathcal{J}^{(l)}_{\text{3d}} - \hat{\mathcal{J}}^{(l)}_{\text{3d}}||_2 \tag{6}$$

$$\mathcal{L}_{\text{j2d}} = ||\Pi(\mathcal{J}^{(l)}_{\text{3d}}) - \hat{\mathcal{J}}^{(l)}_{\text{2d}}||_2 \tag{7}$$

$$\mathcal{L}_{\text{v3d}} = ||V^{(l)}_{\text{3d}} - \hat{V}^{(l)}_{\text{3d}}||_2 \tag{8}$$

$$\mathcal{L}_{\text{v2d}} = ||\Pi(V^{(l)}_{\text{3d}}) - \hat{V}^{(l)}_{\text{2d}}||_2 \tag{9}$$

$$\mathcal{L}_{\theta} = \frac{1}{2}\left(||\theta^{*(l)} - \hat{\theta}^{(l)}||_2 + ||\theta^{(l)} - \hat{\theta}^{(l)}||_2\right) \tag{10}$$

which are calculated for every input frame l. Locations of 3D joints and vertices of the SMPL-X model are denoted by $\mathcal{J}_{\text{3d}}$ and V_{3d}. The hat operator denotes the ground truth and Π is the camera projection operator. We only slightly adjust the joint rotation loss $\mathcal{L}_\theta$ such that it is calculated for the initial $\theta^{*(l)}$ and refined joint rotations $\theta^{(l)}$, which differ only in the ankle rotations.

Architectural details. FootMR consists of six RoPE-based transformer encoder layers. Each attention unit features four attention heads, and the hidden dimension is set to $d_\mathrm{H} = 256$. MLPs have two linear layers with GELU activation. The maximum window size for self-attention is $W = 120$. This corresponds to a temporal window of four seconds when using videos with 30 frames per second.

4. Experiments

Evaluation datasets. To quantitatively evaluate our approach, accurate 3D foot annotations are required. Unfortunately, SMPL annotations provided by popular in-the-wild benchmark datasets such as 3DPW [52] are often not accurate for the feet (see Fig. 2, *right*). Thus, we instead evaluate on the multi-view datasets MOYO [50] and RICH [12]. MOYO [50] contains videos of a Yoga professional performing highly complex poses including extreme foot poses. The subject is captured using a marker-based motion capture system, resulting in highly accurate ground truth SMPL-X fits. RICH [12] is recorded in both indoor and outdoor environments and reconstructs 3D human ground truth bodies using markerless motion capture. Multiple subjects are recorded performing a mixture of daily (*e.g.*, cooking, cleaning, eating) and sporting activities (*e.g.*, lunges, push-ups, burpees). RICH contains significantly less foot motion than MOYO, and due to using markerless motion capture, the annotations are less accurate. MOYO contains only a single subject and all videos are recorded in the same indoor environment. Furthermore, the extreme out-of-domain body poses could lead to degenerated pose predictions, making it difficult to evaluate foot movement in isolation. Due to these limitations we introduce the **MOOF** dataset (complex **MO**vements **O**f the **F**eet). We

record subjects performing simple body movements that involve complex foot motions, such as ankle circles, ankle stretches, and heel–toe walking, and augment these recordings with in-the-wild dance and ballet videos collected online. MOOF comprises 15 subjects (9 female, 6 male) and 41 videos captured at 30 fps. Video durations range from 4 seconds to 37 seconds, totaling 14,589 frames. We use a semi-automatic annotation pipeline to obtain 2D ground truth keypoints for the big toe, small toe and heel. Two example images from MOOF are shown in Fig. 4. More details are provided in the supplementary material.

Evaluation metrics. The accuracy of 3D human motion estimation is typically evaluated by computing the Mean Per Joint Position Error (MPJPE), Procrustes-aligned MPJPE (PA-MPJPE), and the Per Vertex Error (PVE). Because joints or vertices of the feet represent only a small percentage of the entire body, improved foot poses have very little impact on the whole-body metrics. Therefore, we instead compute the following foot-specific 3D metrics: the global Ankle Joint Angle Error (AJAE, in degrees) and the scale-normalized [40] MPJPE for foot keypoints (big toe, small toe, heel), which we refer to as N-MPJPE$_{\text{F}}$ (reported in mm). Foot keypoints are mean-centered per foot before computing the N-MPJPE$_{\text{F}}$. Evaluation of 3D metrics is done in camera coordinates. On the 2D dataset MOOF, we evaluate the 2D image alignment of reprojected foot keypoints using PCK$_{\text{F}}$ (Percentage of Correct foot Keypoints) and N-FKE$_{\text{2d}}$ (Normalized 2D Foot Keypoint Error). PCK$_{\text{F}}$ measures the percentage of predicted foot keypoints with an L2 distance from the ground truth below a specified threshold. We use a threshold of 0.05, corresponding to 5% of the person's bounding box. Because PCK$_{\text{F}}$ also depends on the predicted shape, translation, and body pose, we additionally measure the N-FKE$_{\text{2d}}$. N-FKE$_{\text{2d}}$ is computed by first normalizing the keypoints using the person's bounding box, then mean-centering and scale-aligning them per foot, and finally computing the L2 distance from the ground truth.

4.1. Implementation Details

SMPL-X motion estimator. To evaluate our foot motion refinement method, we couple it with the state-of-the-art 3D human motion recovery method GVHMR [42]. We use GVHMR because it is efficient and jointly trainable with motion capture and video datasets. GVHMR is a temporal transformer-based model that processes several input features including image features and 2D human keypoints. We additionally create a baseline by slightly modifying GVHMR such that three 2D keypoints per foot (big toe, small toe, heel) are appended to the original 17 input body keypoints, and refer to the baseline as GVHMR$_{\text{23j}}$.

Training. We follow the training setting of GVHMR [42] and use the large-scale motion capture dataset AMASS [32] and the video datasets BEDLAM [2], Human3.6M [13], and 3DPW [52] for training. We train FootMR together with GVHMR from scratch for 500 epochs with a batch size of 256, which takes around 30 hours on a single H100 GPU. The AdamW optimizer [30] is used with a learning rate of 2×10^{-4} that is halved after 200 and 350 epochs. Following [42], the losses presented in Sec. 3.2 are weighted with $\lambda_\theta = 1$, $\lambda_{\text{j3d}} = 500$, $\lambda_{\text{j2d}} = 1000$, $\lambda_{\text{v3d}} = 500$, $\lambda_{\text{v2d}} = 1000$, and the length of the training sequences is set to $L = 120$. During training, ground truth global knee rotations are used as input for FootMR, as this slightly improves the convergence rate. Predicted initial global ankle rotations are detached from the computational graph prior to being used as input. We train a second model using the same setting by replacing GVHMR with GVHMR$_{\text{23j}}$.

4.2. Quantitative Evaluation

We compare our approach with several state-of-the-art per-frame and video-based methods and with our straightforward baseline GVHMR$_{\text{23j}}$. For all competitors, we use the officially released model checkpoint to ensure a consistent and fair comparison. No test-time flip augmentation is performed and ground truth focal lengths are not provided during inference. Quantitative results on MOYO, RICH and MOOF are shown in Table 1. Our foot motion refinement method (FootMR) significantly improves foot motion reconstruction of both GVHMR and of our baseline GVHMR$_{\text{23j}}$, outperforming all competitors by a large margin. Notably, compared to the best existing temporal methods, FootMR achieves an error reduction of 30.6% on MOYO (37.3 to 25.9 AJAE) and of 58.1% on MOOF (1.60 to 0.67 N-FKE$_{\text{2d}}$). The same metrics are improved by 16.5% and 52.8% with respect to the best per-frame method [36]. The strong performance on MOYO and MOOF shows that FootMR successfully generalizes to complex foot movements. The reconstruction of everyday activities are also consistently improved as demonstrated on RICH. Note that CameraHMR [36] achieves better full pose reconstructions (PA-MPJPE), primarily due to training on more diverse in-the-wild data with more accurate pseudo-GT annotations. However, its foot metrics, especially on MOYO and MOOF, are significantly worse, showing that the improved pseudo-GT annotations are still not accurate and diverse enough for the feet. The full-body pose metric PA-MPJPE is only shown for reference, as the focus of this work is solely on improving foot pose reconstruction. Note that the ankle joint angle error (AJAE) cannot be evaluated on RICH for methods that predict SMPL parameters, since RICH only has ground truth in SMPL-X format. All other metrics are computed by first converting the SMPL-X meshes to SMPL format using a vertex mapper [2].

Interestingly, our simple baseline GVHMR$_{\text{23j}}$ already outperforms all competitors on all foot metrics. Intuitively,

	Models	MOYO			RICH			MOOF	
		AJAE↓	N-MPJPE$_F$↓	PA-MPJPE↓	AJAE↓	N-MPJPE$_F$↓	PA-MPJPE↓	PCK$_F$ 0.05↑	N-FKE$_{2d}$↓
per-frame	HMR2.0 [8] ICCV'23	42.0	55.7	83.5	-	36.9	57.2	76.1	1.95
per-frame	ReFit [56] ICCV'23	41.8	47.6	79.8	-	30.3	46.7	81.8	1.68
per-frame	TokenHMR [7] CVPR'24	38.0	48.1	56.6	-	27.7	38.7	74.5	1.85
per-frame	CameraHMR [36] 3DV'25	31.0	39.1	45.3	-	24.8	35.0	88.3	1.42
temporal	WHAM [45] CVPR'24	41.7	49.1	68.8	-	28.3	43.7	68.1	1.94
temporal	TRAM [57] ECCV'24	39.6	47.3	63.2	-	26.5	40.5	68.0	1.60
temporal	GVHMR [42] SIGGRAPH'24	37.3	45.9	64.5	18.0	27.3	39.8	76.1	1.70
temporal	GVHMR$_{23j}$ (Baseline)	29.6	37.5	62.7	16.2	24.2	40.1	89.2	1.06
temporal	GVHMR + **FootMR (Ours)**	26.8	32.8	62.2	**15.8**	23.2	40.4	86.9	**0.67**
temporal	GVHMR$_{23j}$ + **FootMR (Ours)**	**25.9**	**32.4**	62.3	**15.8**	**23.0**	39.9	**92.6**	**0.67**

Table 1. **Comparison of foot pose reconstruction** on the MOYO [50], RICH [12], and MOOF datasets. Our foot motion refinement method (FootMR) effectively improves foot motion reconstruction of GVHMR and GVHMR$_{23j}$, outperforming all competitors on all foot-specific metrics by a large margin. Note that improving full-body pose metrics is not the focus of this work; PA-MPJPE (shown in a faint column) is included only for reference. AJAE is in degrees, MPJPE in mm, PCK$_F$ in %, and N-FKE in normalized pixel space.

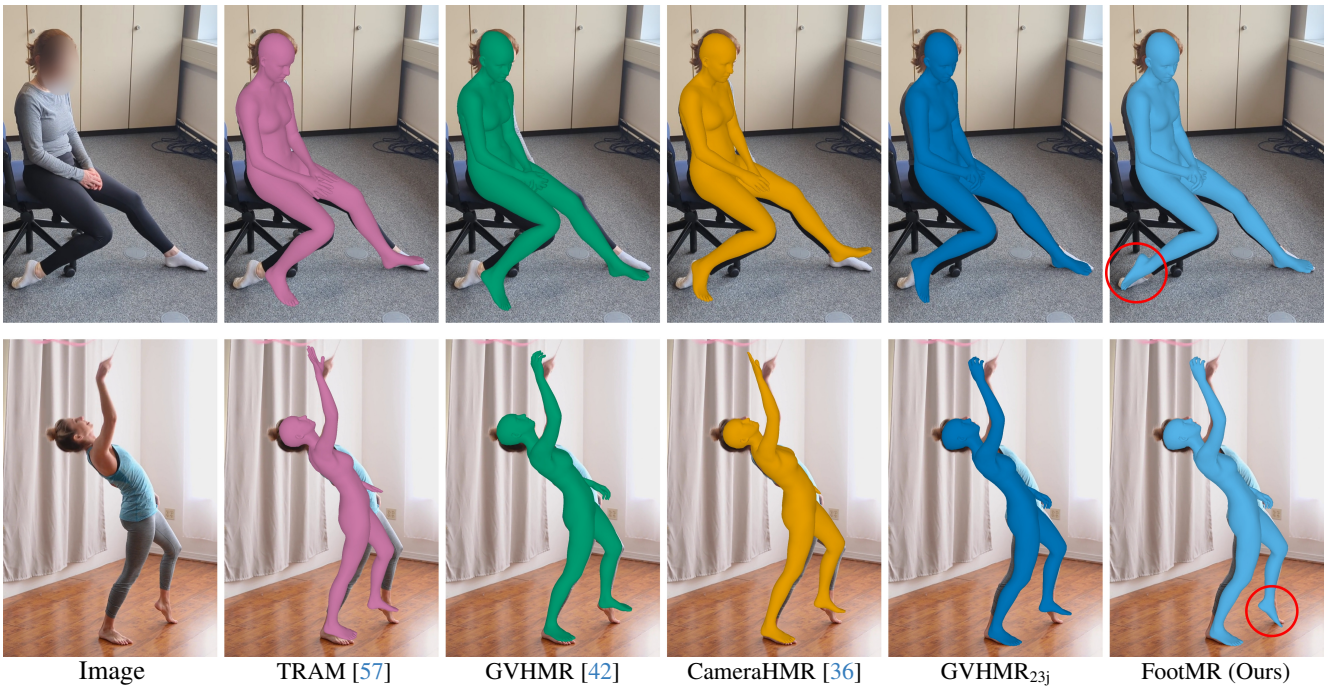

Figure 4. **Qualitative comparison** on the introduced MOOF dataset. FootMR is the only method that is able to accurately reconstruct the extreme foot poses.

by additionally providing the three 2D keypoints per foot, GVHMR$_{23j}$ can fully exploit large-scale motion capture data during training to learn about 3D foot motion. However, unlike FootMR, it still cannot generalize to extreme foot poses, which is evident in the poorer performance on MOOF (N-FKE$_{2d}$ of 1.06 vs. 0.67) and also visually shown in Fig. 4. Combining FootHMR with GVHMR$_{23j}$ achieves slightly better results than combining it with GVHMR. We notice that using the additional 2D foot keypoints leads to GVHMR$_{23j}$ estimating more accurate knee joint rotations. For instance, the knee joint angle error on MOYO improves from 20.6° to 18.9°. This facilitates the ankle joint refinement task and results in better 2D image alignment, as can be seen by the PCK$_F$.

Qualitative comparisons with state-of-the-art competitors are shown in Fig. 4. All methods predict similar body poses, but only FootMR is able to accurately reconstruct the extreme foot poses. For more qualitative results, please refer to our supplementary materials.

4.3. Ablation Study

To analyze the impact of each component of our method, we evaluate multiple variants of FootMR combined with GVHMR$_{23j}$ using the same training and evaluation proto-

Models	MOYO AJAE	RICH AJAE	MOOF N-FKE$_{2d}$
Relative	27.6	16.4	0.78
Global	26.4	19.5	0.73
Residual Relative	28.7	**15.8**	0.90
Residual Global (Ours)	**25.9**	**15.8**	**0.67**
Ours w/o DA	27.3	16.5	0.82

Table 2. **Ablation of output rotation representation.** Estimating residual rotations with respect to initial global ankle rotations from a SMPL-X motion estimator leads to the best results.

Input Joint Rotations				MOYO	RICH	MOOF
Ankle	Knee	Hip	Pelvis	AJAE	AJAE	N-FKE$_{2d}$
×	×	×	×	38.0	36.0	0.65
√	×	×	×	29.1	17.4	**0.61**
√	√	×	×	**25.9**	15.8	0.67
√	√	√	×	26.1	**15.7**	0.71
√	√	√	√	26.0	**15.7**	0.71

Table 3. **Ablation of input joint rotations.** Using global knee and ankle rotations as input for foot motion refinement is important, while additional joints provide no further benefit.

col. We begin by investigating the influence of different rotation representations of the ankle rotations predicted by FootMR. The results are shown in Table 2. The models in the first two rows receive global knee rotations as input and directly estimate the parent-relative or global ankle rotations, instead of residual rotations. When estimating parent-relative rotations, the model cannot generalize well to extreme foot poses, which is reflected in the poorer performance on MOYO and MOOF. The reason could be the limited range of ankle rotations observed during training. Directly estimating global ankle rotations leads to performance degradation on RICH, especially when 2D keypoints are missing due to occlusions. Combining the best of both worlds, FootMR estimates residual rotations with respect to initial global ankle rotations from a SMPL-X motion estimator. This significantly improves robustness to noisy or missing 2D keypoints, while generalizing well to extreme foot poses. It also achieves better results than predicting residual rotations with respect to initial parent-relative ankle rotations. Table 2 additionally shows the impact of the proposed random rotation data augmentation technique, which consistently improves all foot metrics.

Next, we analyze which joint rotations are important as input for refining the ankle rotations. The ablation study is shown in Table 3. A model that directly predicts global ankle rotations only from 2D foot keypoint sequences heavily degenerates when input keypoints are noisy, leading to very poor results on MOYO and RICH. This demonstrates the high ambiguity of the 2D-to-3D mapping. Interestingly, it still manages to achieve good 2D image alignment, showing that 2D foot metrics alone are no reliable indicator to assess the 3D reconstruction performance. Using initial global ankle rotations as input for refinement leads to less ambiguity. Global knee rotations further help to disambiguate the 2D keypoint sequences to 3D rotations mapping. Finally, providing additional joints along the kinematic chain (hip and pelvis) offers no further advantages. More ablation studies are included in the supplementary material.

5. Limitations

FootMR achieves state-of-the-art foot motion reconstruction by accurately capturing the movement of the left and right ankle joint. However, a single ankle joint is not sufficient to model the full range of motion of the foot. Additionally improving the forefoot joint of SMPL-X only helps to a limited extent. Feet on the SMPL-X body are overly simplistic and cannot model movements such as curling of the toes [35]. Future work should explore extending our approach to the articulated foot model SUPR-Foot [35] which contains 13 joints per foot. More dense 2D keypoints will be needed to estimate the additional degrees of freedom.

6. Conclusion

FootMR is a novel approach for improving 3D foot motion reconstruction in markerless monocular human motion capture. It refines initial ankle rotation estimates from an existing 3D human recovery model by leveraging detected 2D foot keypoints. Our experiments show that incorporating knee and initial ankle rotations and predicting only residual ankle rotations effectively resolves ambiguity of lifting the 2D foot keypoint sequences to 3D. By not using images directly as input, FootMR does not suffer from inaccurate foot annotations present in most existing human datasets. It also enables our proposed random root orientation data augmentation which further improves generalization to extreme foot poses. To support evaluation, we introduced MOOF, a new video dataset containing complex foot motion. FootMR combined with the 3D human motion recovery model GVHMR produces accurate and temporally coherent 3D human body and foot motion, outperforming all competitors across diverse benchmarks on various foot-specific metrics. Most notably, FootMR is the only method that is able to accurately reconstruct extreme foot poses not seen in common datasets.

Acknowledgements. This work was supported by the Federal Ministry of Education and Research (BMBF), Germany, under the AI service center KISSKI (grant no. 01IS22093C), the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122), and the European Union within the Horizon Europe research and innovation programme under grant agreement no. 101136006 – XTREME.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. Human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2

[2] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 2, 6

[3] 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*, 2016. 3

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2019. 3

[5] 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 *CVPR*, 2021. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[7] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *CVPR*, 2024. 7

[8] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 3, 7

[9] Peng Guan, Alexander Weiss, Alexandru O. Bãlan, and Michael J. Black. Estimating human shape and pose from a single image. In *ICCV*, 2009. 3

[10] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *CVPR*, 2010. 3

[11] Jaewoo Heo, George Hu, Zeyu Wang, and Serena Yeung-Levy. DeforHMR: Vision transformer with deformable cross-attention for 3d human mesh recovery. In *International Conference on 3D Vision (3DV)*, 2025. 3

[12] 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*, 2022. 2, 5, 7

[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 (PAMI)*, 36(7), 2014. 2, 3, 6

[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. *arXiv preprint arXiv:2303.07399*, 2023. 3

[15] 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 *ECCV*, 2020. 3

[16] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 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 *International Conference on 3D Vision (3DV)*, 2021. 2, 3

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

[19] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 3

[20] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *ECCV*, 2024. 2, 3

[21] Imry Kissos, Lior Fritz, Matan Goldman, Omer Meir, Eduard Oks, and Mark Kliger. Beyond weak perspective for monocular 3d human pose estimation. In *ECCV Workshops*, 2020. 3

[22] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 3

[23] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *ICCV*, 2021. 3

[24] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. PACE: Human and camera motion estimation from in-the-wild videos. In *International Conference on 3D Vision (3DV)*, 2024. 3

[25] 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. 2, 3

[26] Jifeng Li, Ye Yuan, Davis Rempe, Haotian Zhang, Pavlo Molchanov, Cewu Lu, Jan Kautz, and Umar Iqbal. COIN: Control-inpainting diffusion prior for human and camera motion estimation. In *ECCV*, 2024. 3

[27] 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*, 2022. 2, 3, 4

[28] 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 *ECCV*, 2014. 2, 3

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *SIGGRAPH Asia*, 2015. 3

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[31] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. In *ACCV*, 2020. 3

[32] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 6

[33] 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 *International Conference on 3D Vision (3DV)*, 2017. 2

[34] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *CVPR Workshops*, 2022. 2, 3

[35] Ahmed A A Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. SUPR: A sparse unified part-based human body model. In *ECCV*, 2022. 3, 8

[36] Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. In *International Conference on 3D Vision (3DV)*, 2025. 3, 6, 7

[37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 3

[38] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *ECCV*, 2020. 3

[39] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 3

[40] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. *CVPR*, 2018. 6

[41] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *CVPR*, 2023. 3

[42] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia*, 2024. 1, 3, 4, 5, 6, 7

[43] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics*, 2020. 3

[44] Soyong Shin, Zhixiong Li, and Eni Halilaj. Markerless motion tracking with noisy video and imu data. *IEEE Transactions on Biomedical Engineering*, 2023. 5

[45] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *CVPR*, 2024. 3, 5, 7

[46] Leonid Sigal, Alexandru Balan, and Michael Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NeurIPS*, 2007. 3

[47] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 4

[48] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[49] 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*, 2023. 3

[50] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *CVPR*, 2023. 5, 7

[51] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2

[52] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018. 2, 5, 6

[53] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *ICCV*, 2021. 3

[54] Shengze Wang, Jiefeng Li, Tianye Li, Ye Yuan, Henry Fuchs, Koki Nagano, Shalini De Mello, and Michael Stengel. BLADE: Single-view body mesh estimation through accurate depth estimation. In *CVPR*, 2025. 3

[55] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction supplementary material. In *ICCV*, 2023. 3

[56] Yufu Wang and Kostas Daniilidis. ReFit: Recurrent fitting network for 3d human recovery. In *ICCV*, 2023. 7

[57] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3d humans from in-the-wild videos. *ECCV*, 2024. 3, 7

[58] Tom Wehrbein, Bodo Rosenhahn, Iain Matthews, and Carsten Stoll. Personalized 3d human pose and shape refinement. In *ICCV Workshops*, 2023. 3

[59] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Utilizing uncertainty in 2d pose detectors for probabilistic 3d human mesh recovery. In *WACV*, 2025. 3

[60] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *CVPR*, 2022. 3

[61] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3d human shape and articulated pose models. In *CVPR*, 2020. 3

[62] 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 (PAMI)*, 2022. 3

[63] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer foundation model for generic body pose estimation. *arXiv preprint arXiv:2212.04246*, 2022.

[64] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV Workshops*, 2023. 3

[65] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. 3

[66] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics*, 2021. 3

[67] 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*, 2022. 3

[68] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, Lei Yang, and Ziwei Liu. WHAC: World-grounded humans and cameras. In *ECCV*, 2024. 3

[69] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. 3

[70] 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*, 2021. 3

[71] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3, 4

[72] Yue Zhu, Nermin Samet, and David Picard. H3WB: Human3.6m 3d wholebody dataset and benchmark. In *ICCV*, 2023. 3

[73] Li'an Zhuo, Jian Cao, Qi Wang, Bang Zhang, and Liefeng Bo. Towards stable human pose estimation via cross-view fusion and foot stabilization. In *CVPR*, 2023. 3

[74] Yuliang Zou, Jimei Yang, Duygu Ceylan, Jianming Zhang, Federico Perazzi, and Jia-Bin Huang. Reducing footskate in human motion reconstruction with ground contact constraints. In *WACV*, 2020. 3