Decoupling Human and Camera Motion from Videos in the Wild

Vickie Ye   Georgios Pavlakos   Jitendra Malik   Angjoo Kanazawa

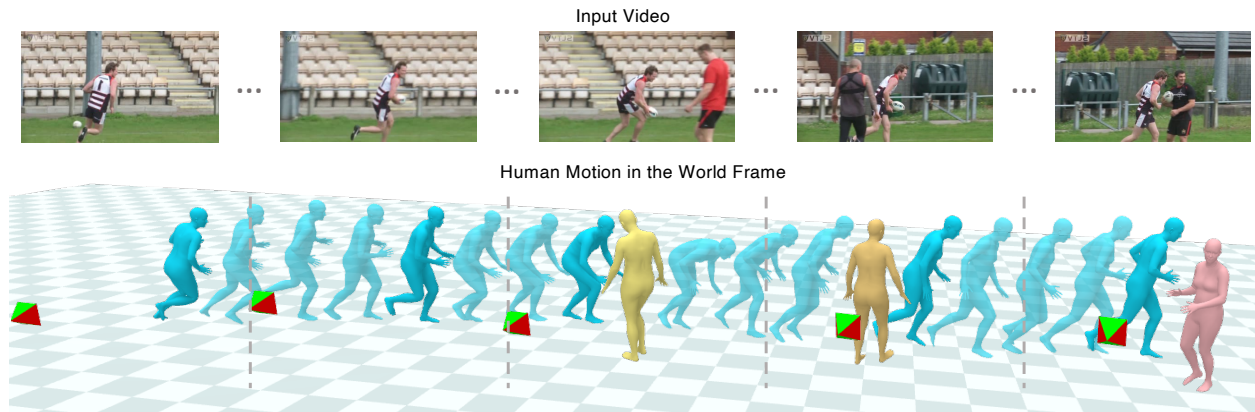University of California, Berkeley

Figure 1. **4D Reconstruction of People from Videos in-the-Wild.** We present **SLAHMR**: **S**imultaneous **L**ocalization **A**nd **H**uman **M**esh **R**ecovery, a method that given a video of moving people (top), recovers the global trajectories of all people and cameras in the world coordinate space (bottom). We combine geometric insights, which determine relative camera motion, with learned human motion priors, which constrain a person's plausible displacement between frames, to position the people and cameras in the shared world frame through time. Our method can recover the global trajectories of all detected people from in-the-wild videos with uncontrolled camera and human motion. Please see the project page to see the full video results.

Abstract

We propose a method to reconstruct global human trajectories from videos in the wild. Our optimization method decouples the camera and human motion, which allows us to place people in the same world coordinate frame. Most existing methods do not model the camera motion; methods that rely on the background pixels to infer 3D human motion usually require a full scene reconstruction, which is often not possible for in-the-wild videos. However, even when existing SLAM systems cannot recover accurate scene reconstructions, the background pixel motion still provides enough signal to constrain the camera motion. We show that relative camera estimates along with data-driven human motion priors can resolve the scene scale ambiguity and recover global human trajectories. Our method robustly recovers the global 3D trajectories of people in challenging in-the-wild videos, such as PoseTrack. We quantify our improvement over existing methods on 3D human dataset Egobody. We further demonstrate that our recovered camera scale allows us to reason about motion of multiple people in a shared coordinate frame, which improves performance of downstream tracking in PoseTrack.

1. Introduction

Consider the video sequence in Figure 1. As human observers, we can clearly perceive that the camera is following the athlete as he runs across the field. However, when this dynamic 3D scene is projected onto 2D images, because the camera tracks the athlete, the athlete appears to be at the center of the camera frame throughout the sequence — *i.e.* the projection only captures the *net* motion of the underlying human and camera trajectory. Thus, if we rely only on the person's 2D motion, as many human video reconstruction methods do, we cannot recover their original trajectory in the world (Figure 1 bottom left). To recover the person's 3D motion in the world (Figure 1 bottom right), we must also reason about how much the camera is moving.

We present an approach that models the camera motion to recover the 3D human motion in the world from videos in the wild. Our system can handle multiple people and reconstructs their motion in the same world coordinate frame, enabling us to capture their spatial relationships. Recovering the underlying human motion and their spatial relationships is a key step towards understanding humans from in-the-wild videos. Tasks, such as autonomous planning in environments with humans [50], or recognition of human interactions with

the environment [68] and other people [21, 37], rely on information about global human trajectories. Current methods that recover global trajectories either require additional sensors, e.g. multiple cameras or depth sensors [11, 51], or dense 3D reconstruction of the environment [12, 31], both of which are only realistic in active or controlled capture settings. Our method acquires these global trajectories from videos in the wild, with no constraints on the capture setup, camera motion, or prior knowledge of the environment. Being able to do this from dynamic cameras is particularly relevant with the emergence of large-scale egocentric video datasets [5, 9, 69].

To do this, given an input RGB video, we first estimate the relative camera motion between frames from the static scene's pixel motion with a SLAM system [56]. At the same time, we estimate the identities and body poses of all detected people with a 3D human tracking system [45]. We use these estimates to initialize the trajectories of the humans and cameras in the shared world frame. We then optimize these global trajectories over multiple stages to be consistent with both the 2D observations in the video and learned priors about how human move in the world [46]. We illustrate our pipeline in Figure 2. Unlike existing works [11, 31], we optimize over human and camera trajectories in the world frame without requiring an accurate 3D reconstruction of the static scene. Because of this, our method operates on videos captured in the wild, a challenging domain for prior methods that require good 3D geometry, since these videos rarely contain camera viewpoints with sufficient baselines for reliable scene reconstruction.

We combine two main insights to enable this optimization. First, even when the scene parallax is insufficient for accurate scene reconstruction, it still allows reasonable estimates of camera motion up to an arbitrary scale factor. In fact, in Figure 2, the recovered scene structure for the input video is a degenerate flat plane, but the relative camera motion still explains the scene parallax between frames. Second, human bodies can move realistically in the world in a small range of ways. Learned priors capture this space of realistic human motion well. We use these insights to parameterize the camera trajectory to be both consistent with the scene parallax and the 2D reprojection of realistic human trajectories in the world. Specifically, we optimize over the scale of camera displacement, using the relative camera estimates, to be consistent with the human displacement. Moreover, when multiple people are present in a video, as is often the case in in-the-wild videos, the motions of all the people further constrains the camera scale, allowing our method to operate on complex videos of people.

We evaluate our approach on EgoBody [69], a new dataset of videos captured with a dynamic (ego-centric) camera with ground truth 3D global human motion trajectory. Our approach achieves significant improvement upon the state-of-the-art method that also tries to recover the human motion without considering the signal provided by the background pixels [63]. We further evaluate our approach on PoseTrack [1], a challenging in-the-wild video dataset originally designed for tracking. To demonstrate the robustness of our approach, we provide the results on all PoseTrack validation sequences on our project page. On top of qualitative evaluation, since there are no 3D ground-truth labels in PoseTrack, we test our approach through an evaluation on the downstream application of tracking. We show that the recovered scaled camera motion trajectory can be directly used in the PHALP system [45] to improve tracking. The scaled camera enables more persistent 3D human registration in the 3D world, which reduces the re-identification mistakes. We provide video results and code at the project page.

2. Related Work

Human Mesh Recovery from a Single Image. In the literature for 3D human mesh reconstruction, most methods operate by recovering the parameters of a parametric human body model, notably SMPL [32] or its follow-up models [38, 40, 49, 60]. The main paradigms are optimization-based, e.g., SMPLify [3] and follow-ups [27, 40, 57], or regression-based, like HMR [19] and follow-ups [10, 25, 66]. For regression approaches in particular, many efforts have focused on increasing the model robustness in a variety of settings [18, 23, 26, 41, 48]. Most of these approaches predict the human mesh in the camera coordinate frame with identity camera. There are recent works, e.g., SPEC [24] and CLIFF [30], that also consider incorporating camera information in the regression pipeline, but only for single frame inference. PHALP [45] is a state-of-the-art method on tracking using the predicted 3D information of people ran on each frame. We use the detected identities and predicted 3D mesh as the initialization and show how it can be improved by incorporating the camera obtained by our approach.

Human Mesh Recovery from Video. Many works extend human mesh recovery approaches on video to recover a smooth plausible human motion. However, these works fail to account for camera motion and do not recover global human trajectories. Regression approaches like HMMR [20], VIBE [22], and follow-ups [4, 33, 41] operate on a bounding box level and only consider the local motion of the person within that bounding box. These approaches are prone to jitter since they are sensitive to the bounding box size. More recently, approaches such as GLAMR [63], D&D [28] and Yu et al. [62], have tried to circumvent the issue of camera motion by recovering plausible global trajectories from the per-frame local human poses. However, relying *only* on local pose is not sufficient for a faithful global trajectory, especially for out-of-distribution poses, and is brittle when local pose cannot be fully observed. As such, [63] struggles
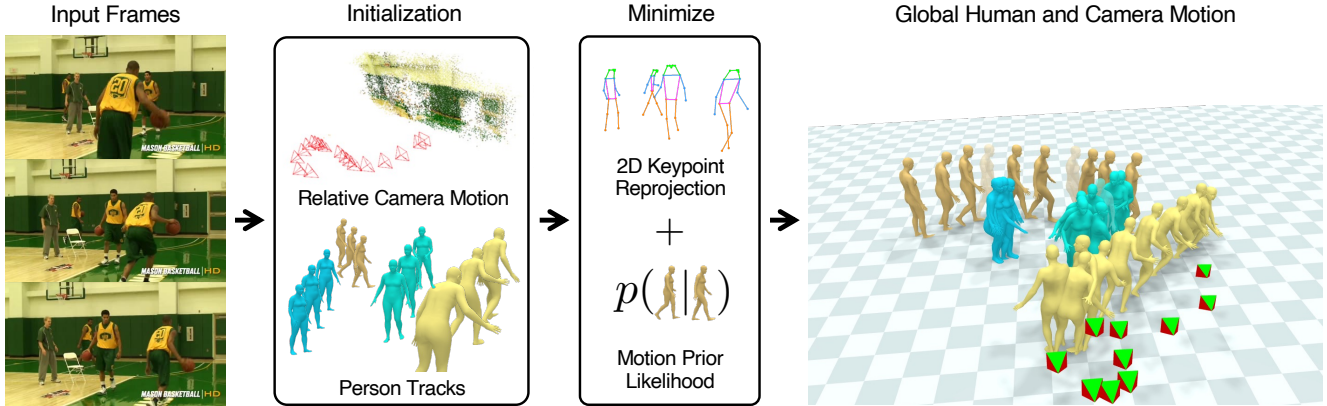
Figure 2. **SLAHMR Pipeline.** Given an input video in-the-Wild with moving camera (left), we first predict the relative camera motion with SfM [56] (middle top). We also recover the unique identities of people across the video and their local 3D poses [45]. These are input into the proposed joint optimization system, which solves for the 4D trajectories of the moving people in the world coordinate frame, as well as the scale and the ground of the world.

on in-the-wild videos, which often have partial occlusions and diverse human actions. Our work explicitly accounts for the camera motion to place the humans in the static scene.

Optimization approaches are similarly limited by the lack of camera awareness. [2, 41] use body pose smoothness priors to recover net human motion over short sequences, ignoring cameras entirely. Recent methods achieve more realistic human motion by modeling human dynamics, through learned priors [46] or physics based priors [7, 43, 47, 52, 53, 59, 64]. These priors are naturally defined in the human coordinate frame, and have thus far been limited to settings where the camera is metrically known, or static. Our approach opens a path in which these methods can be applied to moving cameras.

Other works rely on prior 3D scene information or additional sensors to contextualize human motion. [11, 31, 42] can recover faithful global trajectories when the cameras and dense 3D environment have already been reconstructed. Such reconstructions require observations of the scene from many viewpoints with wide baselines. [11, 31] both rely on reconstructions from actively controlled capture data; [42] rely on television data in which the same set was observed from many different viewpoints. [51] recovers global human trajectories with multiple synchronized cameras, again only realistic for controlled capture settings, or a single static camera. In contrast, our work recovers human trajectories for in-the-wild videos, in which camera motion is uncontrolled, and the scene reconstruction is limited or non-existent. [15] operate on monocular sequences, but the extent of results is limited to a single unoccluded person slowly walking in an indoor studio. We demonstrate our approach on PoseTrack, a complex in-the-wild dataset, which includes videos with a large number of people in various environments.

Human Mesh Recovery for Multiple People. There have been many works that consider the reconstruction of multiple people from single images. Zanfir *et al*. [65] propose an optimization approach, while follow-up work [6, 65, 67] has considered regression solutions. Jiang *et al*. [17] incorporate constraints that encourage the consistency of the multiple people in 3D using a Mask R-CNN [14] type of network, while Sun *et al*. [54, 55] has investigated center-based regression [71]. Mustafa *et al*. [36] consider implicit representations for the multiple person recovery. However, all of the above works operate on a single frame basis. Mehta *et al*. [35] operate on video but they only reconstruct the 3D skeleton and demonstrate results on simpler sequences with a static camera. In contrast we recover the 3D trajectories of multiple people from a moving camera.

3. Method

We take as input a video with T frames of a scene with N people. Our goal is to recover the motion of all detected people in the world coordinate system. We use the SMPL-H model [32, 49] and represent each person i at timestep t via global orientation $\Phi_t^i \in \mathbb{R}^3$, body pose (22 joint angles), $\Theta_t^i \in \mathbb{R}^{22 \times 3}$, shape $\beta^i \in \mathbb{R}^{16}$, shared over all timesteps t, and root translation $\Gamma_t^i \in \mathbb{R}^3$:

$$\mathbf{P}_t^i = \{\Phi_t^i, \Theta_t^i, \beta^i, \Gamma_t^i\}. \tag{1}$$

The SMPL model uses these parameters to generate the mesh vertices $\mathbf{V}_t^i \in \mathbb{R}^{3 \times 6890}$ and joints $\mathbf{J}_t^i \in \mathbb{R}^{3 \times 22}$ of a human body through the differentiable function $\mathcal{M}$:

$$[\mathbf{V}_t^i, \mathbf{J}_t^i] = \mathcal{M}(\Phi_t^i, \Theta_t^i, \beta^i) + \Gamma_t^i. \tag{2}$$

We begin by estimating each person's per-frame pose $\hat{\mathbf{P}}_t^i$ and computing their unique identity track associations

over all frames using state-of-the-art 3D tracking system, PHALP [45]. PHALP estimates poses independently per-frame, and each estimate resides in the camera coordinate frame. In a video, however, a person's motion in the camera coordinates is a composition of the human and camera motion in the world frame, *i.e.*, the net motion:

$$\text{camera motion} \circ \text{human motion} = \text{net motion}. \quad (3)$$

To recover the original world trajectory of each person, we must determine the camera motion contribution to their net perceived motion. We denote the pose in the camera frame as $^{\mathrm{c}}\mathbf{P}_t^i = \{^{\mathrm{c}}\Phi_t^i, \Theta_t^i, \beta^i, {}^{\mathrm{c}}\Gamma_t^i\}$, and the pose estimate in the world frame as $^{\mathrm{w}}\mathbf{P}_t^i = \{^{\mathrm{w}}\Phi_t^i, \Theta_t^i, \beta^i, {}^{\mathrm{w}}\Gamma_t^i\}$; the local pose $\hat{\Theta}_t^i$ and shape $\hat{\beta}^i$ parameters are the same in both.

Our first insight is to use the information in the static scene's pixel motion to compute the relative camera motion between video frames. We use state-of-the-art data-driven SLAM system, DROID-SLAM [56] to estimate the world-to-camera transform at each time t, $\{\hat{R}_t, \hat{T}_t\}$. The camera motion can only be estimated up to an unknown scale of the world, but human bodies and motion can only take on a plausible range of values in the world. In order to ultimately place the people in the world, we must therefore determine α, the relative scale between the displacement of the camera and that of people.

Our second insight is to use priors about human motion in the world to jointly determine the camera scale α and people's global trajectories. In the following sections, we describe the steps we take to initialize and prepare for joint optimization with a data-driven human motion prior. In Section 3.1, we describe how we initialize the multiple people tracks and cameras in the world coordinate frame. In Section 3.2, we describe a smoothing step on the trajectories in the world, to warm-start our joint optimization problem. Finally in Section 3.3, we describe the full optimization of trajectories and camera scale using the human motion prior.

3.1. Initializing people in the world.

We take as input to our joint optimization problem the pose parameters predicted by PHALP in the camera coordinate frame, $\hat{\mathbf{P}}_t^i$, and the world-to-camera transforms estimated with SLAM, $\{\hat{R}_t, \hat{T}_t\}$. We initialize optimization variables $^{\mathrm{w}}\mathbf{P}_t^i$, for all people $i = 0, \ldots, N-1$ and timesteps $t = 0, \ldots, T-1$. The shape β_i and pose Θ_t^i parameters are defined in the human canonical frame, so we use PHALP estimates directly. We initialize the global orientation and root translation in the world coordinate frame using the estimated camera transforms and camera-frame pose parameters.

$$^{\mathrm{w}}\Phi_t^i = R_t^{-1}{}^{\mathrm{c}}\hat{\Phi}_t^i, \qquad {}^{\mathrm{w}}\Gamma_t^i = R_t^{-1}{}^{\mathrm{c}}\hat{\Gamma}_t^i - \alpha R_t^{-1}T_t,$$
$$\beta_i = \hat{\beta}^i, \qquad\qquad \Theta_t^i = \hat{\Theta}_t^i, \qquad (4)$$

where we initialize the camera scale $\alpha = 1$. The joints in the world frame are then expressed as:

$$^{\mathrm{w}}\mathbf{J}_t^i = \mathcal{M}(^{\mathrm{w}}\Phi_t^i, \Theta_t^i, \beta^i) + {}^{\mathrm{w}}\Gamma_t^i. \quad (5)$$

We use the image observations, that is, the detected 2D keypoints $\mathbf{x}_t^i$ and confidences ψ_t^i [61], to define the joint reprojection loss:

$$E_{\text{data}} = \sum_{i=1}^N \sum_{t=1}^T \psi_t^i \rho\big(\Pi_K(R_t \cdot {}^{\mathrm{w}}\mathbf{J}_t^i + \alpha T_t) - \mathbf{x}_t^i\big), \quad (6)$$

where $\Pi_K\left(\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^\top\right) = K\begin{bmatrix} \frac{x_1}{x_3} & \frac{x_2}{x_3} & 1 \end{bmatrix}^\top$ is perspective camera projection with camera intrinsics matrix $K \in \mathbb{R}^{2\times3}$, and ρ is the robust Geman-McClure function [8].

In the first stage of optimization, we align the parameters of the people in the world with the observed 2D keypoints. Because the reprojection loss (6) is very under-constrained, in this stage, we optimize only the global orientation and root translation $\{^{\mathrm{w}}\Phi_t^i, {}^{\mathrm{w}}\Gamma_t^i\}$ of the human pose parameters:

$$\min_{\{\{^{\mathrm{w}}\Phi_t^i, {}^{\mathrm{w}}\Gamma_t^i\}_{t=1}^T\}_{i=1}^N} \lambda_{\text{data}} E_{\text{data}}. \quad (7)$$

We optimize Equation 7 for 30 iterations with $\lambda_{\text{data}} = 0.001$.

3.2. Smoothing trajectories in the world

We next begin optimizing for the camera scale α and the human shape β_i and body pose Θ_t^i parameters. As we begin to update α, we must disambiguate the contribution of camera motion $\{R_t, \alpha T_t\}$ from the contribution of human translation Γ_t^i to the reprojection error of the joints in Equation 6. To do this, we introduce additional priors about how humans move in the world to constrain the displacement of the people to be plausible. We ultimately use an data-driven transition-based human motion prior in our final stage of optimization; to prepare for this, we perform an optimization stage to smooth the transitions between poses in the world trajectories. We use a simple prior of joint smoothness, or minimal kinematic motion:

$$E_{\text{smooth}} = \sum_i^N \sum_t^T \|\mathbf{J}_t^i - \mathbf{J}_{t+1}^i\|^2. \quad (8)$$

We also use priors on shape [3] $E_\beta = \sum_i^N \|\beta^i\|^2$ and pose $E_{\text{pose}} = \sum_{i=1}^N \sum_{t=1}^T \|\zeta_t^i\|^2$, where $\zeta_t^i \in \mathbb{R}^{32}$ is a representation of the body pose parameters Θ_t^i in the latent space of the VPoser model [40]. We add these losses to Equation 7, and optimize over the entire $^{\mathrm{w}}\mathbf{P}_t^i$ and camera scale α:

$$\min_{\alpha, \{\{^{\mathrm{w}}\mathbf{P}_t^i\}_{t=1}^T\}_{i=1}^N} \lambda_{\text{data}} E_{\text{data}} + \lambda_\beta E_\beta + \lambda_{\text{pose}} E_{\text{pose}} + \lambda_{\text{smooth}} E_{\text{smooth}}. \quad (9)$$

We optimize for 60 iterations and use $\lambda_{\text{smooth}} = 5, \lambda_\beta = 0.05, \lambda_{\text{pose}} = 0.04$.

3.3. Incorporating learned human motion priors

We finally introduce a learned motion prior that better captures the distribution of plausible human motions. We use the transition-based motion prior, HuMoR [46], in which the likelihood of a trajectory $\{\mathbf{s}_0, \ldots, \mathbf{s}_T\}$ can be factorized into the likelihoods of transitions between consecutive states, $p_\theta(\mathbf{s}_t|\mathbf{s}_{t-1})$, where $\mathbf{s}_t$ is an augmented state representation used by [46], containing the SMPL pose parameters $\mathbf{P}_t$, as well as additional velocity and joint location predictions. The likelihood of a transition $p_\theta(\mathbf{s}_t|\mathbf{s}_{t-1})$ is modeled by a conditional variational autoencoder (cVAE) as

$$p_\theta(\mathbf{s}_t|\mathbf{s}_{t-1}) = \int_{\mathbf{z}_t} p_\theta(\mathbf{z}_t|\mathbf{s}_{t-1})p_\theta(\mathbf{s}_t|\mathbf{z}_t, \mathbf{s}_{t-1}),$$

where $\mathbf{z}_t \in \mathbb{R}^{48}$ is a latent variable representing the transition between $\mathbf{s}_{t-1}$ and $\mathbf{s}_t$. The conditional prior $p_\theta(\mathbf{z}_t|\mathbf{s}_{t-1})$ is parameterized as a Gaussian distribution with learned mean $\mu_\theta(\mathbf{s}_{t-1})$ and covariance $\sigma_\theta(\mathbf{s}_{t-1})$. We then use this learned prior in an energy term on the latents $\mathbf{z}_t^i$:

$$E_{\text{CVAE}} = -\sum_i^N \sum_t^T \log \mathcal{N}(\mathbf{z}_t^i; \mu_\theta(\mathbf{s}_{t-1}^i), \sigma_\theta(\mathbf{s}_{t-1}^i)). \quad (10)$$

We perform optimization over the initial states $\mathbf{s}_0^i$, the camera scale α, and latent variables $\mathbf{z}_t^i$, for timesteps $t = 1, \ldots, T-1$ and people $i = 0, \ldots, N-1$. We initialize the transition latents $\mathbf{z}_t^i$ from consecutive states $\mathbf{s}_{t-1}^i$ and $\mathbf{s}_t^i$ with the pre-trained HuMoR encoder μ_ϕ, and use the HuMoR decoder Δ_θ to recursively roll out state $\mathbf{s}_t^i$ from the previous state $\mathbf{s}_{t-1}^i$ and current latent $\mathbf{z}_t^i$:

$$\mathbf{z}_t^i = \mu_\phi(\mathbf{s}_t^i, \mathbf{s}_{t-1}^i), \quad \mathbf{s}_t^i = \mathbf{s}_{t-1}^i + \Delta_\theta(\mathbf{z}_t^i, \mathbf{s}_{t-1}^i). \quad (11)$$

We recover the entire trajectories $(\mathbf{s}_0^i, \ldots, \mathbf{s}_T^i)$ for all people i by autoregressively rolling out the initial states $\mathbf{s}_0^i$ with the latents $\mathbf{z}_t^i$ initialized in Eq. 11. We also carry over additional losses E_{stab} from [46] to regularize the predicted velocity and joint location components of $\mathbf{s}_t^i$ to be physically plausible and consistent with the pose parameter components of $\mathbf{s}_t^i$; please see [46] for more details. We denote all prior optimization terms as $E_{\text{prior}} = \lambda_{\text{CVAE}}E_{\text{CVAE}} + \lambda_{\text{stab}}E_{\text{stab}}$.

Following [46], we also optimize for the ground plane of the scene $g \in \mathbb{R}^3$, and use the decoder to predict the probability of ground contact $c(j) \in [0, 1]$ for joints j. We enforce a zero velocity prior on joints that are likely to be in contact with the ground g to prevent unrealistic foot-skate:

$$E_{\text{skate}} = \sum_i^N \sum_t^T \sum_j^J c_t^i(j)\|\mathbf{J}_t^i(j) - \mathbf{J}_{t+1}^i(j)\|, \quad (12)$$

while also encouraging their distance from the ground to be less than a threshold δ:

$$E_{\text{con}} = \sum_i^N \sum_t^T \sum_j^J c_t^i(j) \max(d(\mathbf{J}_t^i(j), g) - \delta, 0). \quad (13)$$

Here, $d(\mathbf{p}, g)$ defines the distance between point $\mathbf{p} \in \mathbb{R}^3$ and the plane $g \in \mathbb{R}^3$, and we optimize g as a free variable shared across all people and timesteps. We denote these constraints as $E_{\text{env}} = \lambda_{\text{skate}}E_{\text{skate}} + \lambda_{\text{con}}E_{\text{con}}$.

Our optimization problem for this stage is then

$$\min_{\substack{\alpha, g, \{\mathbf{s}_0^i\}_{i=1}^N, \\ \{\{\mathbf{z}_t^i\}_{t=1}^T\}_{i=1}^N}} \begin{aligned} &\lambda_{\text{data}}E_{\text{data}} + \lambda_\beta E_\beta + \lambda_{\text{pose}}E_{\text{pose}} \\ &\quad + E_{\text{prior}} + E_{\text{env}}. \end{aligned} \quad (14)$$

We optimize Equation 14 with an incrementally increasing horizon, increasing T in chunks of 10: $H = 10\tau$, $\tau = 1, \ldots, \lceil \frac{T}{10} \rceil$. We optimize $\{\mathbf{z}_0, \ldots, \mathbf{z}_H\}$ adaptively, rolling out the trajectory by 10 more frames each time the loss decreases less than a threshold γ, for a minimum of 5 iterations and maximum 20 iterations. We use $\lambda_{\text{CVAE}} = 0.075$, $\lambda_{\text{skate}} = 100$, and $\lambda_{\text{con}} = 10$. We perform all optimization with PyTorch [39] using the L-BFGS algorithm with learning rate 1.

3.4. Implementation details

Missing observations: Our approach fills in missing information for every person caused by poor detection and/or occlusion. To initialize the person variables for these frames, *i.e.*, $\Phi_0^i, \Theta_t^i, \Gamma_t^i$, we interpolate Φ_t^i, Θ_t^i and on $SO(3)$ and Γ_t^i in $\mathbb{R}^3$. While no data term is available for these missing frames, motion priors along with the camera motion provide additional signals to these parameters.

Handling multiple people in-the-wild: Although simple in concept, reasoning about multiple people at once in the already big optimization problem is a challenge, particularly since in videos in-the-wild, not all people appear at the same timestamp. People can enter the video at any frame, leave and come back again. Our implementation is designed to handle these cases well. We also use an improved version of PHALP with a stronger detector [29], which we refer to as PHALP+. Please see the appendix for more details. Code is available at the project page.

4. Experimental Results

We demonstrate quantitatively and qualitatively that our approach effectively reconstructs human trajectories in the world. We also demonstrate quantitatively that the camera scale we recover can be used to improve people tracking in videos. We encourage viewing additional video results on the project page.

Datasets. Datasets typically used for evaluation in the 3D human pose literature generally only provide videos captured with a static camera (*e.g.*, Human3.6M [16], MPI-INF-3DHP [34], MuPoTS-3D [35], PROX [13]). 3DPW [58] is a dataset captured with moving cameras, and includes indoor and outdoor sequences of people in natural environments. However, as is also discussed by previous work [63, 70], it only provides 3D pose ground truth in the local frame of the

Method	W-MPJPE↓	WA-MPJPE↓	Acc Error↓
Full system	**141.1**	**101.2**	**25.78**
w/o last stage	234.0	152.9	275.9
PHALP$^+$ w/ SfM	253.6	150.3	302.1
PHALP$^+$ [45]	387.8	204.9	307.6

Table 1. **Ablation of the proposed system on EgoBody [69].** Removing any of the proposed components has a significant effect on the final performance of the system. We report W-MPJPE and WA-MPJPE in mm, and Acc Error in mm/s^2. Note the significant difference motion prior and scale makes in the acceleration error.

Method	Egobody	3DPW†	3DPW*
Full system	79.13	**62.60**	**55.86**
w/o last stage	88.18	64.50	59.19
PHALP$^+$ [45]	**72.16**	64.68	56.70

Table 2. **Comparison of PA-MPJPE for Egobody and 3DPW.** All errors in mm. 3DPW† uses the detected PHALP tracks that best match the ground truth track, the result if the system is run out of the box. 3DPW* uses the ground truth person tracks and is most comparable to existing evaluation protocols.

person, and it is not possible to evaluate the global motion of the person. We use only 3DPW to perform ablations on the reconstructed local pose using our method.

The most relevant dataset that is captured with dynamic cameras and provides ground truth 3D pose in the global frame is the recently introduced **EgoBody dataset** [69]. Ego-Body is captured with a head-mounted camera on an *interactor*, who sees and interacts with a second *interactee*. The camera moves as the interactor moves, and the ground truth 3D poses of the interactee are recorded in the world frame. Because videos are recorded from a head-mounted camera, Egobody videos have heavy body truncations, with the interactee often only visible from chest or waist up.

We also demonstrate our approach on the PoseTrack dataset [1]. PoseTrack is an extremely challenging in-the-wild dataset originally designed for tracking. It spans a wide variety of activities, involving many people with heavy occlusion and interaction. We use PoseTrack to qualitatively demonstrate the robustness of our method, and show many results in Figure 3 and in the Sup. Video. Because there is no 3D ground truth on PoseTrack, we perform quantitative evaluation through the downstream task of tracking. We show that reasoning about the tracks with the scaled camera trajectory recovered by our approach, can boost its state-of-the-art performance on PoseTrack.

Evaluation metrics: We report a variety of metrics, with a focus on metrics that compute the error on the *world* coordinate frame. World PA Trajectory - MPJPE (WA-MPJPE) reports MPJPE [16] after aligning the *entire trajectories* of

Method	W-MPJPE↓	WA-MPJPE↓	Acc Err↓	PA-MPJPE↓
PHALP$^+$ [45]	387.8	204.9	307.6	**72.16**
VIBE [22]	500.4	259.5	524.2	100.5
VIBE-opt [22]	453.2	246.0	481.1	100.4
GLAMR [63]	416.1	239.0	173.5	114.3
SLAHMR	**141.1**	**101.2**	**25.78**	79.13

Table 3. **Comparison with the state of the art on EgoBody dataset [69].** We compare our approach with a variety of state-of-the-art methods for human mesh recovery. GLAMR is the only other approach that attempts to recover the global human motion trajectory, but does so from only local pose transitions, without considering scene pixel motion. Our approach obtains significant improvement in the world trajectory metrics, as well as in the acceleration error.

the prediction and ground truth with Procrustes Alignment. World PA First - MPJPE (W-MPJPE) reports the MPJPE after aligning the *first frames* of the prediction and the ground truth. PA-MPJPE reports the MPJPE error after aligning *every frame* of the prediction and the ground truth. We also report Acceleration Error computed as the difference between the magnitude of the acceleration vector at each joint. Please see the Sup. Mat. for more details on the evaluation protocol. For tracking, we report identity switch metrics; other commonly used tracking metrics measure quality of association and detection, but we use the same detection and association protocol in all baselines, so we omit those.

4.1. Egobody results

To demonstrate the effect of the different components of our system, we first perform an ablation study reporting results in Table 1. We use the metrics presented earlier and we discuss different settings. We start with the result of the full system and remove some key components. We report performance metrics of (i) our method without the last stage of optimization, *i.e.*, without the motion prior and scale ("w/o last stage"), (ii) our method before optimization in the world, *i.e.*, only the PHALP$^+$ results with the estimated cameras, and (iii) the basic results of PHALP$^+$ in the camera frame, without estimated cameras at all. We report metrics on the reconstructed trajectories in the world (W-MPJPE, WA-MPJPE, Acc Error) in Table 1.

For completeness, we also report the PA-MPJPE, which is common in the literature, for the same ablations on both the Egobody and 3DPW datasets. Because 3DPW annotates up to two people's poses for the captured sequences, we evaluate two variants of our method. 3DPW† uses our full system's pre-processing: PHALP$^+$ to detect, track, and estimate people's initial 3D local poses. 3DPW* uses each person's ground-truth tracks, and runs PHALP$^+$'s 3D pose estimation only. We note that 3DPW*, *i.e.*, using the ground truth tracks, is most similar to the current evaluation practices on 3DPW. We report the results in Table 2. We see that
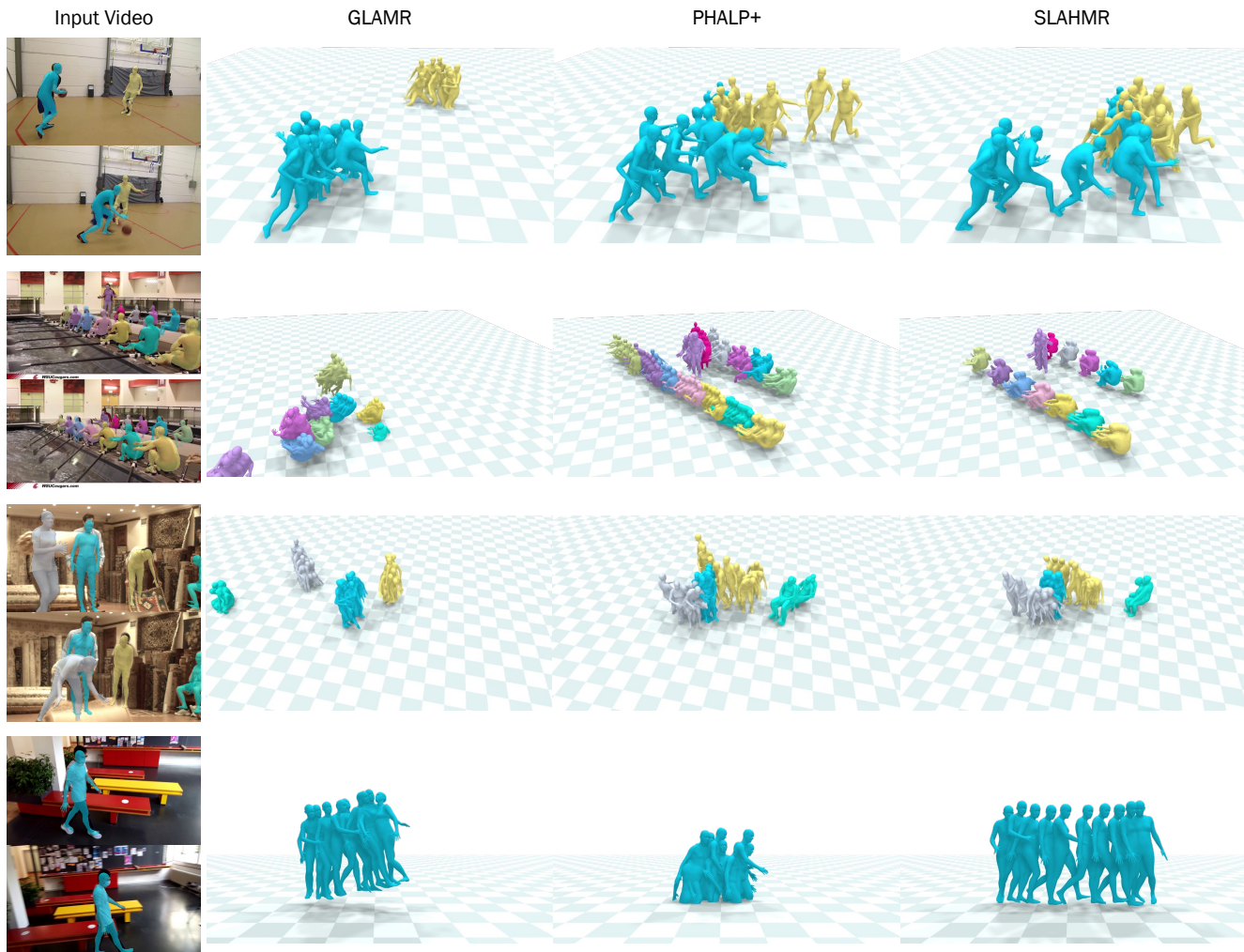
Figure 3. **Qualitative results of the proposed approach.** The first three rows are from PoseTrack [1] and the last row is from EgoBody [69]. The columns compare results on three approaches: GLAMR [63], PHALP$^+$ [45] which is the input to our system, and SLAHMR. We visualize the top-down view of the recovered motion trajectory across the entire frames. Note that GLAMR struggles to recover a plausible global trajectory (first and second row) and also struggles on a sequence with close to static camera input (third row). Our optimization improves upon the inputs in reducing the jitter and recovering plausible motion trajectory with more plausible depth relationship between the people. Please the project page for the results in video format.

in Egobody, in which the subject is often heavily truncated, the prediction method PHALP$^+$ achieves better performance. In other words, further optimization with truncated observations can reduce performance; this is a known issue for mesh recovery methods [18, 23, 41]. However, in 3DPW, in which the subjects are more fully visible, our method improves upon the initial predictions from PHALP$^+$. Ultimately, PA-MPJPE only captures local pose accuracy, and cannot describe the global attributes of the full trajectories.

We also compare our approach with a series of state-of-the-art methods for human mesh recovery in Table 3. The closest to our system is GLAMR [63], which also estimates 3D body reconstructions in the world frame. As we see

in Table 3, we comfortably outperform GLAMR. Because GLAMR computes the world trajectory based on local pose estimates alone, it is especially sensitive to the extreme truncation in Egobodoy videos. In contrast to GLAMR, we leverage the relative camera motion to achieve significantly better reconstruction results, globally and locally.

We also compare against human mesh recovery baselines that compute the motion in the camera frame only. We include state-of-the-art baselines for a) single frame mesh regression (PHALP [45]), b) temporal mesh regression (VIBE [22]), and c) temporal mesh optimization (VIBE-opt [22]). Our method outperforms all baselines in world-level metrics, and is second best in local pose metrics.
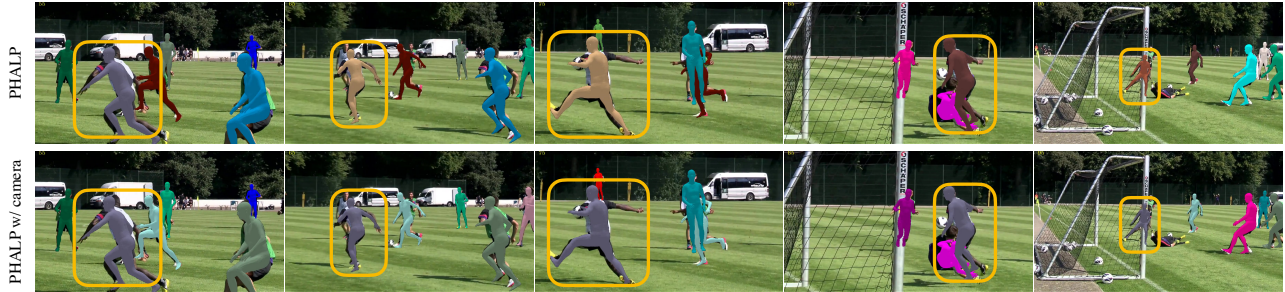
Figure 4. **Effect of the estimated camera for tracking.** We integrate our predicted camera with the PHALP tracking system [45] and we show the effect on the PoseTrack dataset [1]. Explicit modeling of the camera makes PHALP tracking robust to abrupt camera motions. Notice how out-of-the-box PHALP (top row) leads to multiple identity switches, while when we integrate the camera (bottom row), we can achieve consistent tracking over the duration of the video (*i.e.*, the different people maintain the color indicating their identity).

4.2. Posetrack results

For the PoseTrack dataset, we show qualitative results in Figure 3, comparing against GLAMR and PHALP$^+$ (an input to our method). We also provide qualitative reconstructions of all Posetrack validation videos in the supplemental video at the project page. We highly encourage seeing the qualitative results in video to appreciate the improvements in world trajectories. We see that our method recovers smoother trajectories that are more consistent with the dynamic scene in the input videos. We see that GLAMR struggles to plausibly position multiple people in the same world frame.

We also demonstrate a downstream application of our method, by providing helpful camera information to the PHALP tracking method [45]. In brief, PHALP uses an estimate of the 3D person location in the *camera* frame for tracking. We posit that tracking is better done in the world coordinate frame, as it will be invariant to camera motion. We provide to PHALP the recovered camera motion from our approach, *i.e.*, relative cameras from [56] with the scale factor α from our optimization to place the people in the world coordinate frame. We make minimal adaptation to the PHALP algorithm to demonstrate the effect of camera information, however there is a potential for even more improvement. Please see the supplemental for more details. The rest of the tracking procedure operates as in PHALP. We report the results in Table 4. We report both PHALP and PHALP$^+$, which uses a better detection system [29], along with two variants using additional camera information: (i) using the cameras of [56], without rescaling to the scale of human motion, and (ii) also using the recovered scale α from our optimization. We see that using out-of-the-box cameras from [56] does not change performance. In contrast, using the recovered scale from our approach makes a significant improvement to the ID switch metric. This observation indicates our method recovers a more accurate scale, and shows the benefit it can have in the challenging tracking scenario. In Figure 4, we also demonstrate an example of the better behavior we achieve with our tracking.

Method	IDs↓
T3DP [44]	655
PHALP [45]	541
PHALP$^+$	450
PHALP$^+$ + [56] cams	446 (-0.8%)
PHALP$^+$ + [56] cams + α	420 (-6.7%)

Table 4. **Effect of tracking with camera information on PoseTrack [1].** We start with PHALP$^+$, which is PHALP [45] using a more accurate detector [29]. Directly using cameras from [56] yield a small difference in ID switches. However, using the scale α we recover, yields a significant improvement, and highlights the benefit of integrating reliable camera information for tracking.

5. Discussion

We propose a method for recovering human motion trajectories in the world coordinate frame from challenging videos with moving cameras. Our approach leverages relative camera estimates from scene pixel motion to optimize trajectories jointly with learned human motion priors for all people in the scene. This allows us to outperform state-of-the-art methods on the Egobody dataset and generate plausible trajectories for scenes with multiple people and challenging camera motions, as demonstrated by our experiments on the PoseTrack dataset.

While our system unlocks many new sources of human data, many problems remain to be addressed. In-the-wild videos often have ill-posed multiview geometry, such as predominantly rotational camera motion or co-linear motion between humans and cameras. Our method can recover inconsistent trajectories in these cases. Please see the supplemental video for examples. An exciting avenue for future work would be to incorporate human motion priors to also constrain and update the camera and scene reconstruction.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 2, 6, 7, 8

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, 2019. 3

[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. 2, 4

[4] 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. 2

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 2

[6] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. REMIPS: Physically consistent 3D reconstruction of multiple interacting people under weak supervision. *NeurIPS*, 2021. 3

[7] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3D human pose from monocular video. In *CVPR*, 2022. 3

[8] Stuart Geman and Donald E McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 4:5–21, 1987. 4

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2

[10] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, 2019. 2

[11] 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*, 2021. 2, 3

[12] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, pages 2282–2292, Oct. 2019. 2

[13] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *ICCV*, 2019. 5

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 3

[15] Dorian Henning, Tristan Laidlow, and Stefan Leutenegger. BodySLAM: Joint camera localisation, mapping, and human motion tracking. In *ECCV*, 2022. 3

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 2013. 5, 6

[17] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 3

[18] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation. In *3DV*, 2021. 2, 7

[19] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[20] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 2

[21] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013. 2

[22] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2, 6, 7

[23] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 2, 7

[24] 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. 2

[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

[26] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 2

[27] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 2

[28] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&D: Learning human dynamics from dynamic camera. In *ECCV*, 2022. 2

[29] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 5, 8

[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*, 2022. 2

[31] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M Rehg, and Siyu Tang. 4D human body capture from egocentric video via 3D scene grounding. In *3DV*, 2021. 2, 3

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[33] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3D human motion estimation via motion compression and refinement. In *ACCV*, 2020. 2

[34] 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 *3DV*, 2017. 5

[35] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *Acm Transactions On Graphics (TOG)*, 39(4):82–1, 2020. 3, 5

[36] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. Multi-person implicit reconstruction from a single image. In *CVPR*, 2021. 3

[37] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *CVPRR*, 2020. 2

[38] Ahmed AA Osman, Timo Bolkart, and Michael J Black. STAR: Sparse trained articulated human body regressor. In *ECCV*, 2020. 2

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[40] 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*, 2019. 2, 4

[41] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *CVPR*, 2022. 2, 3, 7

[42] Georgios Pavlakos, Ethan Weber, Matthew Tancik, and Angjoo Kanazawa. The one where they reconstructed 3D humans and environments in TV shows. In *ECCV*, 2022. 3

[43] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. SFV: Reinforcement learning of physical skills from videos. *ACM Transactions On Graphics (TOG)*, 37(6):1–14, 2018. 3

[44] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3D representations. In *NeurIPS*, 2021. 8

[45] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location and pose. In *CVPR*, 2022. 2, 3, 4, 6, 7, 8

[46] 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. 2, 3, 5

[47] 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

[48] Chris Rockwell and David F. Fouhey. Full-body awareness from partial observations. In *ECCV*, 2020. 2

[49] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017. 2, 3

[50] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020. 1

[51] Nitin Saini, Chun-hao P. Huang, Michael J. Black, and Aamir Ahmad. Smartmocap: Joint estimation of human and camera motion using uncalibrated rgb cameras, 2022. 2, 3

[52] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3D human motion capture with physical awareness. *ACM Transactions on Graphics*, 40(4), aug 2021. 3

[53] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. PhysCap: Physically plausible monocular 3D motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 3

[54] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *ICCV*, 2021. 3

[55] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3D people in depth. In *CVPR*, 2022. 3

[56] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. *NeurIPS*, 2021. 2, 3, 4, 8

[57] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-NDF: Modeling human pose manifolds with neural distance fields. In *ECCV*, 2022. 2

[58] 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. 5

[59] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *ICCV*, 2021. 3

[60] 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. 2

[61] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 4

[62] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021. 2

[63] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. 2, 5, 6, 7

[64] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3D human pose estimation. In *CVPR*, 2021. 3

[65] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3D sensing of multiple people in natural images. *NIPS*, 2018. 3

[66] 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. 2

[67] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, 2021. 3

[68] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020. 2

[69] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Marc Pollefeys, Federica Bogo, and Siyu Tang. EgoBody: Human body shape, motion and social interactions from head-mounted devices. In *ECCV*, 2021. 2, 6, 7

[70] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4D human body capture in 3D scenes. In *ICCV*, 2021. 5

[71] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 3