PACE: Human and Camera Motion Estimation from in-the-wild Videos

Muhammed Kocabas[1,2,3] Ye Yuan[1] Pavlo Molchanov[1] Yunrong Guo[1] Michael J. Black[2]
Otmar Hilliges[3] Jan Kautz[1] Umar Iqbal[1]

[1]NVIDIA [2] Max Planck Institute for Intelligent Systems, Tubingen, Germany [3]ETH Zurich, Switzerland

https://nvlabs.github.io/PACE/

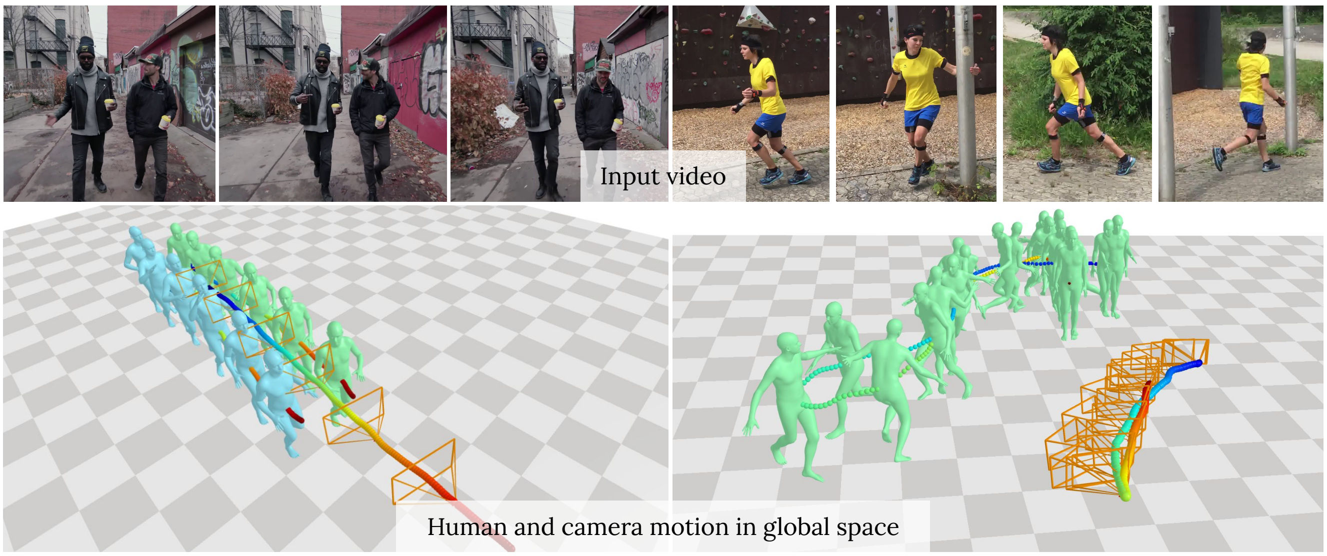


Figure 1. **Human and camera motion reconstruction from in-the-wild videos:** given a video of multiple people, PACE is able to reconstruct the motions of all humans and the camera in a *coherent* global space. To achieve this, we leverage the benefits of both camera localization methods and human motion priors, exploiting the complementary nature of these approaches, *i.e*., dynamic foreground motion *vs*. static background features, to address each other's limitations.

Abstract

We present a method to estimate human motion in a global scene from moving cameras. This is a highly challenging task due to the coupling of human and camera motions in the video. To address this problem, we propose a joint optimization framework that disentangles human and camera motions using both foreground human motion priors and background scene features. Unlike existing methods that use SLAM as initialization, we propose to tightly integrate SLAM and human motion priors in an optimization that is inspired by bundle adjustment. Specifically, we optimize human and camera motions to match both the observed human pose and scene features. This design combines the strengths of SLAM and motion priors, which leads to significant improvements in human and camera motion estimation. We additionally introduce a motion prior that is suitable for batch optimization, making our approach significantly more efficient than existing approaches. Finally, we propose a novel synthetic dataset that enables evaluating camera motion in addition to human motion from dynamic videos. Experiments on the synthetic and real-world RICH datasets demonstrate that our approach substantially outperforms prior art in recovering both human and camera motions.

The work was done during Muhammed's internship at NVIDIA.

1. Introduction

Jointly estimating global human and camera motion from dynamic RGB videos is an important problem with numerous applications in areas such as robotics, sports and mixed reality. However, it is a very challenging task because the observed human and camera motions in the video are entangled. Estimating human motion by itself from videos is highly under-constrained since subject and camera motion are interchangeable. Analogously, camera motion estimation is more challenging in dynamic scenes due to spurious correspondences. Finally, pure monocular approaches can only estimate camera trajectories up to scale.

There are only a few works that address the problem of global pose estimation [55, 106, 107]. These methods leverage the insight that the global human root trajectory is correlated with the local body movements; *e.g*., observing a running motion is indicative of forward motion. Hence, they suggest that global root trajectories can be estimated by exploiting learned motion priors [107] or by enforcing physics-based constraints on the reconstructed human motion [55, 106]. While this idea can help to estimate global human trajectories, motion priors or physical constraints are not enough to fully resolve the ambiguity in the mapping from local motion to global trajectories, especially under root rotations. Others utilize SLAM methods (*e.g*., COLMAP) to estimate camera poses [63, 105], then keep the camera poses fixed and estimate the global scale. However, in-the-wild videos often contain moving objects which can degrade the camera pose localization and subsequently affect the human motion estimates.

In this paper, we propose a novel approach, called PACE (Person And Camera Estimation), to tackle the above problems. We formulate the problem as a global optimization and jointly optimize human and camera motions, leveraging a bundle adjustment objective to match both human pose and background scene features. In this way, the SLAM algorithm uses mostly static scene features, that do not correspond to human motion. Simultaneously, the human motion prior helps correct inaccurate camera trajectories that are incompatible with the local body movements, and informs about the global scale based on human motion statistics. We show that this formulation provides robustness to inaccurate initial human or camera motion estimates.

A further contribution lies in the human motion prior itself. Commonly used human priors *e.g*., HuMoR [86] are typically autoregressive and become prohibitively slow when incorporated in a per-frame optimization, in particular for long motion sequences. In this work, we show that neural motion field (NeMF [30]) can be used to design a parallel motion prior that drastically improves computational efficiency. We divide the entire sequence into overlapping clips and maximize the likelihood of the human motion under the prior. This results in a significantly more efficient implementation without compromising reconstruction quality. Notably, the parallel motion prior allows the runtime of PACE to grow sub-linearly w.r.t. the sequence length in contrast to the linear rate in prior work.

Since it is difficult to obtain ground-truth human and camera poses for in-the-wild videos, we also propose a new synthetic dataset for benchmarking human and camera motion estimation from dynamic videos called the Human and Camera Motion (HCM) dataset. It is the first dataset that provides ground-truth human and camera motion information for this task. We will make the dataset publicly available to facilitate research in this direction.

We evaluate PACE on two datasets: the newly proposed synthetic HCM dataset and the RICH dataset [34], which contains a moving camera with ground truth 3D human pose and shape. Results show that our method substantially outperforms state-of-the-art (SOTA) approaches in accurately recovering human motions from dynamic cameras. Notably, our method also significantly improves camera motion estimation over SOTA SLAM algorithms for this task, which demonstrates the advantage of our global optimization framework. Additionally, we conduct extensive ablation studies to validate the impact of various design choices on performance.

In summary, our contributions are as follows:

- We present a novel approach for precise global human and camera motion estimation from dynamic cameras, which tightly integrates human motion priors and SLAM into a unified optimization framework that leverages both human pose and scene information.
- We propose a parallel motion prior optimization scheme, which significantly improves efficiency without sacrificing accuracy, and allows the runtime to grow sub-linearly w.r.t. the sequence length.
- We introduce HCM, a synthetic dataset for benchmarking global human and camera motion estimation.
- Our method outperforms the SOTA methods significantly in recovering both human and camera motions, achieving 52% and 74% improvements respectively, which fully demonstrate the synergy of our unified approach.

2. Related Work

Camera-Space Human Pose Estimation. Due to the difficulty in monocular depth estimation, most existing methods estimate human poses in the coordinate frame centered around the pelvis of the human body [3, 7, 10–12, 24, 41, 43, 44, 48, 50–54, 62, 65, 75–78, 80, 86, 88, 91, 92, 94, 100, 103, 113, 117, 122, 128]. These methods adopt an orthographic camera projection model and ignore the absolute 3D translation of the person with respect to the camera. To overcome this limitation, recent methods estimate human meshes in the camera coordinates [37, 40, 58, 63, 82, 85, 89, 101, 114, 116, 118]. Some methods use an optimization framework to recover the absolute translation of the person [70–72, 87, 115] or exploit various scene constraints to improve depth prediction [99, 114]. Others employ physics-based constraints to ensure the physical plausibility of the estimated poses [13, 21, 38, 89, 101, 112], use limb-length constraints [36] or approximate depth using the bounding box size [40, 74, 118]. Several approaches employ inverse kinematics to estimate human meshes with absolute translations in the camera coordinates [37, 58]. Heatmap-based representations have also been used to directly predict the absolute depths of multiple people [19, 93, 126]. A few methods learn to also predict the camera parameters from the image, which are used for absolute pose regression in the camera coordinates [49, 60, 116]. While these methods achieve impressive results for camera-relative pose estimation, they fail to decouple human and camera motions from dynamic videos, and therefore cannot recover global human trajectories as our method does.

Global Human Pose Estimation. The majority of current methods for estimating 3D poses in world coordinates rely on synchronized, calibrated, and static multi-view capture setups [6, 14–16, 18, 33, 42, 84, 85, 123, 124, 127]. Huang *et al*. [8] use uncalibrated cameras but still assume time synchronization and static camera setups. Hasler *et al*. [27] handle unsynchronized moving cameras but assume multi-view input and rely on an audio stream for synchronization. Recently, Dong *et al*. [17] proposed to recover 3D poses from unaligned internet videos of different actors performing the same activity from unknown cameras, assuming that multiple viewpoints of the same pose are available in the videos. Luvizon *et al*. [67] estimate the global human poses of multiple people using the scene point cloud for static cameras. In contrast, our approach estimates human meshes in global coordinates from *monocular* videos recorded with dynamic cameras. Several methods rely on additional IMU sensors or pre-scanned environments to recover global human motions [25, 79, 98], which is impractical for large-scale adoption. Another line of work has recently focused on estimating accurate human-scene interaction [29, 34, 66, 106]. Recent work uses human motion priors [107] and physics-based constraints [55, 106] to decouple human and camera motions but does not consider background scene features, which limits performance on in-the-wild videos. Liu *et al*. [63] obtain global human pose using SLAM and convert the pose from the camera to global coordinates. BodySLAM [31] uses features of both humans and scenes, but it only demonstrates results of a single unoccluded person slowly walking in an indoor scene. Along this line, a recent work [105] obtains initial camera trajectories with SLAM and optimizes the scale of the camera trajectories using a human motion prior [86]. In contrast, our approach tightly integrates SLAM and human motion priors into a joint optimization framework, where the entire SLAM camera trajectories (not only scale) are optimized jointly to match observed human pose and background scene features. This not only leads to more accurate human trajectory estimation but also improves full camera trajectory estimation over SLAM significantly, which has not been achieved by prior work. Additionally, our parallel motion optimization scheme also makes our approach substantially (50 times) faster than [105] for a sequence of 1000 frames. Our parallel scheme also allows PACE's time cost to grow sub-linearly w.r.t. sequence length in contrast to the linear rate of [105].

Human Motion Prior. There has been a significant amount of research on 3D human dynamics for various tasks, including motion prediction and synthesis [4, 5, 9, 20, 23, 28, 39, 61, 69, 81, 83, 97, 104, 108–110]. Recently, human pose estimation methods have started to incorporate learned human motion priors to help resolve pose ambiguity [48, 86, 121]. Motion-infilling approaches have also been proposed to generate complete motions from partially observed motions [26, 32, 45, 46]. Diffusion models [90] have also been used as priors for motion synthesis and infilling [35, 96, 111, 119]. Recently, He *et al*. [30] proposes the neural motion field (NeMF), which expresses human motion as a time-conditioned continuous function and demonstrates superior motion synthesis performance. Our approach extends NeMF by leveraging it as a motion prior for human pose estimation. Additionally, our proposed parallel motion optimization scheme enables efficient optimization of human motions.

3. Method

The input to PACE is an in-the-wild RGB video $\mathbf{I}=\{\mathbf{I}_1, \cdots, \mathbf{I}_T\}$ with T frames captured by a moving camera. Our goal is to estimate both the camera motion and the motion of all visible people in the video in a global world coordinate system. The camera motion $\{\mathbf{R}_t, \mathbf{T}_t\}_{t=1}^{T}$ consists of the camera rotation $\mathbf{R}_t \in \mathbb{R}^{3\times3}$ and translation $\mathbf{T}_t \in \mathbb{R}^3$ for every timestep t in the video. The global motion $\mathbf{Q}^i=\{Q_t^i=\{\Phi_t^i, \tau_t^i, \theta_t^i, \beta^i\}\}_{t=s^i}^{e^i}$ for person i consists of the global translation $\tau_t^i \in \mathbb{R}^3$, global orientation $\Phi_t^i \in \mathbb{R}^{3\times3}$, and the body pose parameters $\theta_t^i \in \mathbb{R}^{23\times3}$ for all time steps $t \in \{s^i \cdots e^i\}$, where s^i and e^i correspond to the first and last frame in which person i is visible. The body shape parameters β^i are shared across all time steps. We use the SMPL body model [64] to obtain the articulated body meshes $\mathbf{V}^i=\{V_t^i\}_{t=s^i}^{e^i}$ from $\mathbf{Q}^i$. Specifically, SMPL consists of a linear function $\mathcal{M}(\Phi, \tau, \theta, \beta)$ that maps the body motion $Q_t^i=(\Phi_t^i, \tau_t^i, \theta_t^i, \beta_t^i)$ to a triangulated body mesh $V_t^i \in \mathbb{R}^{6890\times3}$ with 6890 vertices. In the rest of this paper, we drop the superscript i from all variables for brevity but always assume the visibility of multiple people.

Our key insight is to harness the complementary properties of SLAM and human motion priors. The human motion prior can be used to explain foreground human motion, which typically is dynamic and therefore has been treated as unwanted noise in existing SLAM algorithms. Leveraging the motion prior in a joint optimization regularizes the camera trajectories to be in agreement with plausible human motion and provides information about the global scale. On the other hand, SLAM leverages mostly static background

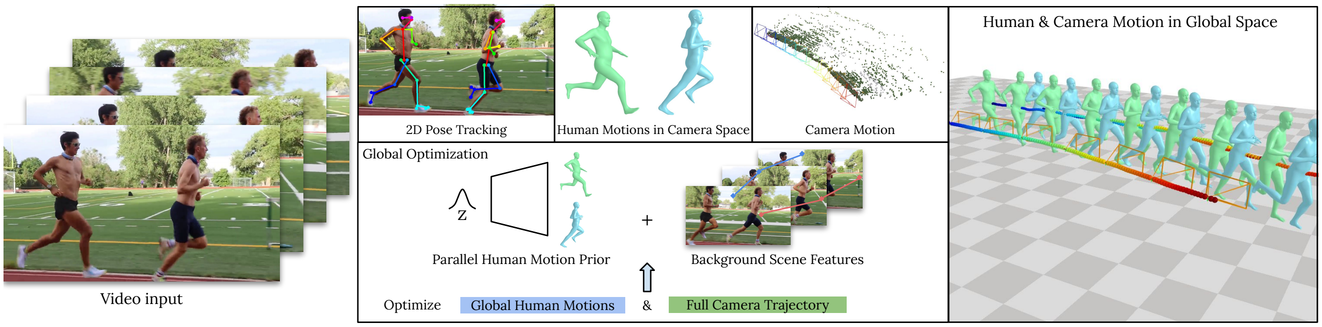


Figure 2. **PACE overview.** Given a video with dynamic human and camera motions, we first use off-the-shelf methods to obtain initial 2D human pose, 3D human motion, and camera motions. We propose a unified optimization framework that optimizes the global human motions and full camera trajectories to reduce 2D pose errors, increase motion likelihood under human motion prior, and match background features. The final output is coherent human and camera motion in global space.

features, which provide information about the camera motion and can be leveraged to resolve ambiguity in the motion space of the human motion priors.

We introduce a novel unified framework, illustrated in Fig. 2, that simultaneously recovers the camera and human motion using a joint optimization objective (Sec. 3.3). Since this is a highly ill-posed problem, we exploit data-driven models to initialize our objective (Sec. 3.1) and use human motion priors to constrain the solution space (Sec. 3.2).

3.1. Initialization

We start by obtaining bounding box sequences for all visible subjects using an off-the-shelf multi-object tracking and re-identification algorithm [125]. We then estimate body pose information for each detected bounding box using the state-of-the-art method HybrIK [58]. HybrIK provides body poses in the camera coordinate frame which we represent as $\hat{Q}^c_t{=}(\hat{\Phi}^c_t, \hat{\tau}^c_t, \hat{\theta}_t, \hat{\beta}_t)$. The super-script c corresponds to the camera coordinate frame. Note that the local body pose θ_t and shape β_t are agnostic to camera motion. For videos recorded with dynamic cameras, the estimated translation $\hat{\tau}^c_t$ and root orientation $\hat{\Phi}^c_t$ must be transformed from camera coordinates to a consistent world coordinate frame. This requires knowledge of the per-frame camera-to-world transforms $\{R_t, T_t\}^T_{t=1}$. For this, we leverage a data-driven SLAM method, namely DROID-SLAM [95], which uses the information of the static scene to estimate per-frame camera-to-world transforms $\{\hat{R}_t, \hat{T}_t\}^T_{t=1}$. SLAM methods, however, provide camera translations $\hat{T}_t$ up to scale. Hence, at this stage, we only use the camera rotation information to obtain a person's root orientation in the world coordinate frame: $\hat{\Phi}_t = \hat{R}_t^{-1}\hat{\Phi}^c_t$. We then use a neural network similar to [30, 56] to estimate the initial global root translations $\{\hat{\tau}_t\}^e_{t=s}$ from the local pose parameters $\{\hat{\Phi}_t, \hat{\theta}_t\}^e_{t=s}$. We use a single value for shape parameters β for each person that we initialize with the average of the per-frame estimates from HybrIK i.e., $\hat{\beta}{=}\frac{\sum_{t=s}^{e}\hat{\beta}_t}{e-s}$. This forms our initial estimate of the global human motion $\hat{Q}{=}\{\hat{Q}_t{=}(\hat{\Phi}_t, \hat{\tau}_t, \hat{\theta}_t, \hat{\beta})\}^e_{t=s}$ in the world coordinate frame. In the remainder of this paper, our goal is to refine these initial estimates via human motion priors and the background scene features, while recovering accurate global camera trajectories.

3.2. Human Motion Prior

Our goal is to develop a human motion prior that ensures that the estimated human motion is plausible and also helps constrain the solution space during joint optimization of human and camera motion. For this, we use a variational autoencoder (VAE) [47], which learns a latent representation $\mathbf{z}$ of human motion and regularizes the distribution of the latent code to be a normal distribution. We want the decoder $\mathcal{D}$ of the VAE to be non-autoregressive for faster sampling while not sacrificing accuracy. This is important because we want to use the motion prior in an iterative optimization, and auto-regressive motion priors (e.g., HuMoR [86]) are prohibitively slow when processing large motion sequences. In contrast, a non-autoregressive decoder can be evaluated for the entire sequence in parallel. To this end, we adopt a Neural Motion Field (NeMF) [30] based decoder to represent body motion as a continuous vector field of body poses via a NeRF-style MLP [73]. In Sec. 3.3, we show that NeMF can be extended to a parallel motion prior that enables efficient optimization. We follow [30] and only model the local body motion via the prior. Specifically, $\mathcal{D}$ is an MLP that takes the latent codes $\{\mathbf{z}_\Phi, \mathbf{z}_\theta\}$ and a time step t as input and produces the orientation $\hat{\Phi}_t$, local body pose $\hat{\theta}_t$, and joint contacts $\hat{\kappa}_t$ for a given time step:

$$\mathcal{D} : (t, \mathbf{z}_\Phi, \mathbf{z}_\theta) \rightarrow (\hat{\Phi}_t, \hat{\theta}_t, \hat{\kappa}_t), \tag{1}$$

where $\mathbf{z}_\Phi$ and $\mathbf{z}_\theta$ control the root orientation Φ and the local body pose θ of the person, respectively. For a given pair of $\mathbf{z}_\Phi$ and $\mathbf{z}_\theta$ the entire sequence can be sampled in parallel by simply varying the values of t. To incorporate the motion priors during global optimization, we optimize the latent codes $\{\mathbf{z}_\Phi, \mathbf{z}_\theta\}$ instead of directly optimizing the local body motion $\{\Phi_t, \theta_t\}_{t=s}^{e}$. We initialize the latent codes using the pre-trained encoders of the VAE; *i.e.*, $z_\Phi = \mathcal{E}_\Phi(\{\Phi\}_{t=s}^{e})$ and $z_\theta = \mathcal{E}_\theta(\{\theta\}_{t=s}^{e})$. We refer to [30] for training details.

Global Translation Estimation. We use a fully convolutional network to generate the global translation τ_t^i of the root joint, based on the local joint positions, velocities, rotations, and angular velocities as inputs. All quantities can be computed from joint rotations. Our approach, which is similar to [57, 129], takes into account the fact that the subject's global translation is conditioned on its local poses. In order to avoid any ambiguity in the output, we predict the velocity $\dot{\tau}_t$ rather than τ_t directly, and then integrate the velocity using the forward Euler method to obtain $\tau_{t+1} = \tau_t + \dot{\tau}_t \Delta t$. We also predict the height of the root joint using the same convolutional network to prevent any cumulative errors that could cause the subject to float above or sink into the ground.

Since changing the latent codes $\{\mathbf{z}_\Phi, \mathbf{z}_\theta\}$ also impacts the global translations τ_t, for simplicity, we refer to the mapping from latent codes to global human motion as

$$\mathcal{P} : (t, \mathbf{z}_\Phi, \mathbf{z}_\theta) \rightarrow (\hat{\Phi}_t, \hat{\theta}_t, \hat{\tau}_t). \tag{2}$$

3.3. Global Optimization

Here we detail the proposed optimization formulation for the joint reconstruction of global human and camera motion. Our goal is to optimize the latent code $\mathbf{z} = \{\mathbf{z}_\Phi, \mathbf{z}_\theta\}$ and camera-to-world transforms $\{R_t, sT_t\}$ with correct scale s. Note that SLAM methods assume the camera at the first frame ($t = 0$) to be at the origin. To align all coordinate frames, we also optimize the camera height h_0 and orientation R_0 for the first frame. More specifically, we optimize the following objective function:

$$\min_{\substack{\beta, \mathbf{z} \\ s, h_0, R_0, \{R_t, T_t\}_{t=1}^{T}}} E_{\text{body}} + E_{\text{scene}} + E_{\text{camera}}, \tag{3}$$

where

$$\begin{aligned} E_{\text{body}} &= E_{\text{2D}} + E_\beta + E_{\text{pose}} + E^{\text{b}}_{\text{smooth}} \\ &\quad + E_{\text{VAE}} + E_{\text{consist}}, \\ E_{\text{scene}} &= E_{\text{contact}} + E_{\text{height}}, \\ E_{\text{camera}} &= E_{\text{PCL}} + E^{\text{c}}_{\text{smooth}}. \end{aligned}$$

The error term E_{body} ensures that the reconstructed human motion is plausible and agrees with the image evidence. E_{2D} measures the 2D reprojection error between the estimated 3D motion and 2D body joints $\mathbf{x}_t$ obtained using a state-of-the-art 2D joint detector [102]:

$$E_{\text{2D}} = \sum_{i=1}^{N} \sum_{t=s_i}^{e_i} \omega_t \zeta(\Pi(R_0 R_t J_t^i + sT_t + \begin{bmatrix} 0 \\ 0 \\ h_0 \end{bmatrix}) - \mathbf{x}_t^i). \tag{4}$$

Here ω_t are the body joint detection confidences, ζ is the robust Geman-McClure function [22], Π corresponds to perspective projection using the known camera intrinsic matrix K, and J_t^i corresponds to 3D body joints that are obtained from the SMPL body mesh via a pre-trained regressor $\mathcal{W}$:

$$J_t^i = \mathcal{W}(\mathcal{M}(\mathcal{P}(\mathbf{z}, t), \beta_t^i)). \tag{5}$$

The error term E_{pose} penalizes large deviations of the local body pose $\hat{\theta}_t$ from the HybrIK predictions, E_β is prior over body shapes [43], and E_{VAE} a motion prior loss defined as:

$$\begin{aligned} E_{\text{VAE}} = -\sum_{i}^{N} &\log \mathcal{N}(\mathbf{z}_\Phi^i; \mu_\Phi(\{\Phi_t^i\}), \sigma_\Phi(\{\Phi_t^i\})) + \\ &\log \mathcal{N}(\mathbf{z}_\theta^i; \mu_\theta(\{\theta_t^i\}), \sigma_\theta(\{\theta_t^i\})). \end{aligned} \tag{6}$$

The term E_{contact} encourages zero velocities for joints that are predicted to be in contact $\hat{\kappa}_t$ with the ground plane:

$$E_{\text{contact}} = \sum_{i=1}^{N} \sum_{t=s_i}^{e^i} \hat{\kappa}_t^i ||J_t^i - J_{t-1}^i||^2, \tag{7}$$

where $\hat{\kappa}_t^i \in \mathbb{R}^{24}$ is the contact probability output from the motion prior decoder $\mathcal{D}$ for each joint. E_{height} prevents in-contact joints from being far away from the ground plane:

$$E_{\text{height}} = \hat{\kappa}_t^i \max(|J_t^i| - \delta, 0). \tag{8}$$

The ground plane is kept fixed and assumed to be xy-plane aligned with $+z$-axis as the up direction. This parameterization allows us to optimize all variables in this consistent coordinate frame without the need to optimize an additional ground plane equation.

The error term E_{camera} in Eq. (3) ensures that the reconstructed camera motion is smooth and consistent with the static scene motion. Since DROID-SLAM is trained on videos with static scenes only, its estimates can be noisy due to the dynamic humans present in our target videos. Hence, we propose to use the point cloud recovered by SLAM as a direct constraint in our optimization, instead of directly relying on the camera predictions. To ensure that the points on dynamic humans do not influence camera reconstruction, we remove all points that lie inside the person bounding boxes. The term E_{PCL} then computes the re-projection error of the pruned point cloud similar to Eq. (4). The term $E^{\text{b}}_{\text{smooth}}$ ensures that the optimized parameters are temporally smooth.

We empirically chose the weights of different error terms in our objective and provide more details in the appendix (Table 5).

Parallel Motion Optimization. Our specific choice of human motion prior, NeMF [30], allows us to design a parallel motion prior that is suitable for batch optimization, which

Stages	Opt. Variables	Loss Functions	Description
Stage-1	s, h_0, R_0, β	$E_{2D} + E_\beta$	camera traj. transform
Stage-2	$s, h_0, R_0, \beta, \mathbf{z}_\Phi$	$E_{\text{body}} + E_{\text{scene}}$	+ global human orientation
Stage-3	$s, h_0, R_0, \beta, \mathbf{z}_\Phi, \mathbf{z}_\phi$	$E_{\text{body}} + E_{\text{scene}}$	+ local body pose
Stage-4	$\beta, \mathbf{z}_\Phi, \mathbf{z}_\phi, R_t, T_t$	$E_{\text{body}} + E_{\text{scene}} + E_{\text{camera}}$	+ full camera trajectory

Table 1. Optimization stages.

significantly enhances the efficiency of our approach. Concretely, we split a motion sequence into overlapping windows of T=128 frames. We use 16 overlapping frames to help reduce jitter and discontinuities across windows. Dividing motions into overlapping windows also allows the latent codes of the prior to model a fixed length of motion. Since our motion prior is non-autoregressive, we can optimize all windows in parallel. To ensure smooth transitions between clips we additionally compute a batch consistency term E_{consist}, defined as the ℓ_2 distance between 3D joints J_t^i of overlapping frames.

Multi-Stage Optimization. The task of reasoning about the camera and human motion from a video is inherently ill-posed, as optimizing both camera motion $R_t.T_t$ and motion prior latent codes $\{\mathbf{z}_\Phi, \mathbf{z}_\theta\}$ simultaneously can result in local minima. To address this challenge, we adopt a multi-stage optimization pipeline, with different parameters optimized in different stages to avoid bad minima. After obtaining initial camera motion results from SLAM and human motion results from the motion prior, the optimization process is carried out in four stages, as outlined in Table 1. In Stage-1, we optimize only the first frame camera parameters (R_0, h_0), camera scale s, and the subjects' body shape β based on the initial camera and human motion. In Stage-2, we incorporate the global orientation latent code $\mathbf{z}_\Phi$ to jointly adjust the subjects' global orientation and camera. In Stage-3, we optimize the local body motion $\mathbf{z}_\theta$ as well. Finally, in Stage-4, we jointly optimize the full camera trajectory along with $\mathbf{z}_\Phi$ and $\mathbf{z}_\theta$. Each stage is run for 500 steps. The λ coefficients used for each objective term can be found in the appendix (Table 5).

Occlussion Handling. Our approach offers a natural solution for occlusions due to subjects in the scene. We achieve this by excluding error terms for occluded frames during optimization and solely optimize the latent codes $\{\mathbf{z}_\Phi, \mathbf{z}_\theta\}$ for visible frames. After optimization, we sample motions from the motion prior to infill the missing poses which will be consistent with their visible neighbors.

4. Experiments

We design our experiments to answer the following questions: (1) Can our unified approach, PACE, achieve SOTA human motion estimation performance for dynamic videos? (2) Can PACE improve camera motion estimation of a SOTA SLAM method? (3) What are the critical components in PACE that significantly impact performance?

Figure 3. Some examples of our proposed HCM dataset. (animated figure, see in Adobe Acrobat).

4.1. Datasets and Metrics

HCM Synthetic Dataset. Currently, available datasets that provide dynamic videos (*e.g.*, [63, 98]) for evaluating human pose and shape estimation have been primarily focused on evaluating the accuracy of local body estimation while neglecting the importance of global human motion estimation. Furthermore, evaluation datasets for simultaneous localization and mapping (SLAM) algorithms do not feature humans and do not provide human motion information. As such, there is a need to create a comprehensive dataset that provides accurate labels for global human and camera motion. To address this need, we have created the HCM (Human and Camera Motion) dataset, which enables the evaluation of both human and camera motion. We use the characters from the RenderPeople [1] dataset and animate them in the scenes obtained from Unreal Engine marketplace [2]. We obtain motion capture (MoCap) clips from the AMASS dataset [68]. For camera trajectory, we designed heuristics to replicate typical camera movements observed in everyday videos and professional movies. Final images were rendered using NVIDIA Omniverse. Additional information regarding the data generation process can be found in the appendix (Sec. A.3). Some example sequences can be seen in Fig 3.

RICH Dataset. The RICH dataset [34] was collected using a total of 7 static and one moving camera. While the ground truth poses are available for the persons and static cameras, the ground truth poses of the moving camera are not available. As such, we only assess the performance of global human motion estimation using this dataset.

Metrics. We report various metrics for both human and camera motion, with an emphasis on those that compute the error in world coordinates. Regarding human motion evaluation, the W-MPJPE metric is used to report MPJPE after aligning the first frames of the predicted and ground truth data. The WA-MPJPE metric is used to report MPJPE after aligning the entire trajectories of the predicted and ground truth data using Procrustes Alignment. Additionally, the

Methods		Human Motion Estimation				Camera Motion Estimation		
	W-MPJPE ↓	**WA-MPJPE ↓**	**W-RJE ↓**	**PA-MPJPE ↓**	**ACCEL ↓**	**ATE ↓**	**ATE-S ↓**	**CAM ACCEL ↓**
Initialization	1116.3	650.0	1083.1	67.6	54.3	155.8	1670.7	17.1
Stage-1 (cam. traj. transform)	1116.3	650.0	1083.1	67.6	54.3	155.8	643.0	17.1
Stage-2 (+ global human orientation)	937.0	488.2	901.9	67.6	54.6	155.8	504.9	18.1
Stage-3 (+ local body pose)	904.5	478.2	877.9	66.6	17.6	155.8	501.3	17.3
Stage-4 (+ full cam. traj.) w/o E_{PCL}	870.1	487.4	844.1	67.6	53.2	166.4	505.0	**15.4**
Stage 2-4	978.6	566.2	939.7	89.9	16.1	164.0	550.6	19.7
Stage 3-4	953.2	490.0	923.7	68.8	18.4	160.1	523.2	17.1
HybrIK [58] + SLAM [95]	1137.3	780.3	1100.9	67.6	51.3	155.8	1670.7	17.1
GLAMR [107]	1977.6	653.8	1958.0	86.0	33.4	1295.2	1714.6	282.9
SLAHMR [105]	888.9	483.5	862.2	69.9	**14.9**	155.8	506.5	17.6
PACE (Ours)	**861.2**	**478.3**	**839.5**	**65.3**	16.7	**137.5**	**459.7**	16.2

Table 2. State-of-the-art comparison and ablation studies on the HCM dataset.

Methods	W-MPJPE ↓				
WA-MPJPE ↓	W-RJE ↓	PA-MPJPE ↓	ACCEL ↓		
HybrIK + SLAM	1073.1	404.4	1066.2	46.7	20.2
GLAMR	653.7	365.1	646.6	79.9	107.7
SLAHMR	571.6	323.7	400.5	52.5	9.4
PACE	**380.0**	**197.2**	**370.8**	**49.3**	**8.8**

Table 3. State of the art results on RICH dataset

PA-MPJPE metric is employed to report the MPJPE error after aligning every frame of the predicted and ground truth data. We also include an ACCEL metric that measures the joint acceleration difference between ground-truth and predicted human motion. For camera motion evaluation, we follow SLAM methods and report the average translation error (ATE) after rigidly aligning the camera trajectories, the average translation error without scale alignment (ATE-S), and the CAM ACCEL camera acceleration error. The ATE-S metric provides a more accurate reflection of inaccuracies in the captured scale of the scene.

4.2. Comparison with State-of-the-Art Methods

Human Motion Estimation. We compare PACE with the following baselines on the HCM and RICH datasets: GLAMR [107], SLAHMR [105], SOTA global human and camera estimation approaches; HybrIK [58] + SLAM, which estimates the camera motions using DROID-SLAM [95] and then transforms the human motion estimated by HybrIK from camera to world space. As observed in Tables 2 and 3, PACE outperforms GLAMR, SLAHMR and HybrIK in human motion estimation significantly. In particular, PACE drastically reduces the global pose errors, *i.e.*, decreasing W-MPJPE by 24% and WA-MPJPE 27% on the HCM dataset, and reducing W-MPJPE by 40% and WA-MPJPE by 52% on the RICH dataset. PACE can also recover accurate local human pose, as indicated by better PA-MPJPE on HCM and competitive PA-MPJPE on RICH. Additionally, PACE estimates much smoother motion by reducing the acceleration error (ACCEL) by 50% on HCM and 56% on RICH.

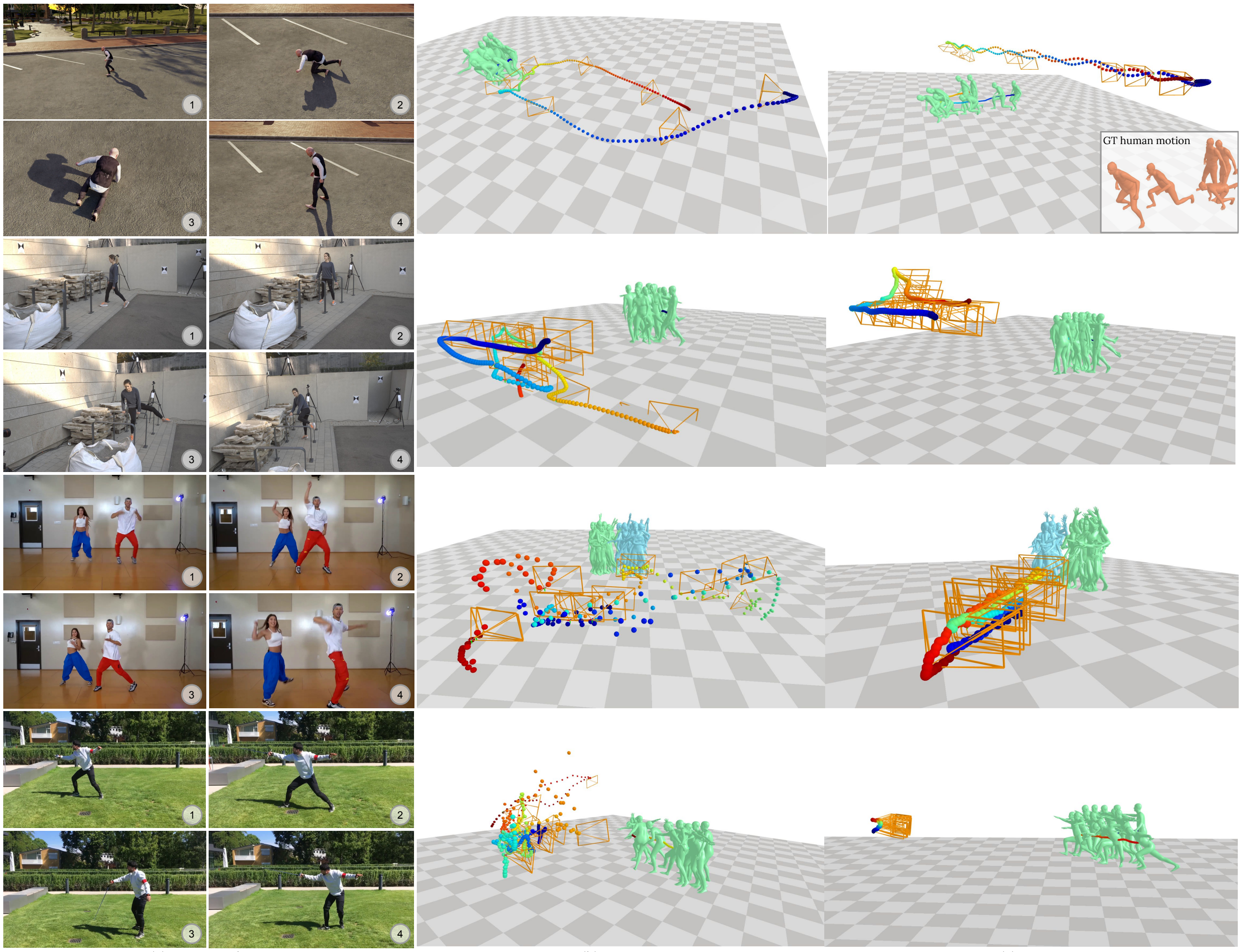
Camera Motion Estimation. The new HCM dataset provides ground-truth camera trajectories that allow us to benchmark camera motion estimation. Table 2 shows that PACE substantially improves the camera motion estimated by a SOTA SLAM algorithm, DROID-SLAM. Specifically, PACE reduces the camera translation error metric, ATE, by 12% with scale alignment and 74% without scale alignment. The above results show that our unified optimization approach can improve both human and camera motion estimation significantly, which answers the first two questions raised at the beginning of this section.

Qualititative Comparison. We also provide qualitative results to visualize the estimated human and camera motions in Fig. 4. Please also refer to the project page for more qualitative results.

Runtime. It is worth noting that the runtime of our optimization framework increases *sub-linearly* w.r.t. sequence length since we can optimize multiple chunks of the motion sequence simultaneously thanks to the parallel motion prior. On average, we can process sequences that are 1000 frames long in less than eight minutes. Notably, SLAHMR [105] reports a runtime of 40 minutes for 100 frames, and this increases linearly with sequence length.

4.3. Ablation Study

We also conduct extensive ablation studies to investigate the effect of each optimization stage and important designs. As shown in Table 2, we compare the performance of PACE after each optimization stage: from stage-1 to stage-4, we gradually add variables to the global optimization – camera trajectory transformation, global human orientation, local body pose, and full camera trajectory (see Sec. 3.3) We observe that gradually adding additional variables to the optimization improves the human motion estimation results. We also try combining stages 2-4 and stages 3-4 to show the importance of multi-stage optimization. Combining these stages drops the performance compared to our 4-stage approach. We also compare PACE against the variant not using the point cloud loss (Stage-4 w/o E_{PCL}) in Table 2. We find that both human and camera motion estimation performance deteriorates when we do not use the point cloud loss,



(a) Input videos (b) GLAMR (c) PACE

Figure 4. **Qualitative results** on HCM (row 1), RICH (row 2), and in-the-wild videos (rows 3 & 4). PACE can estimate more accurate human and camera motion than the SOTA, GLAMR [107], for both datasets and in-the-wild videos.

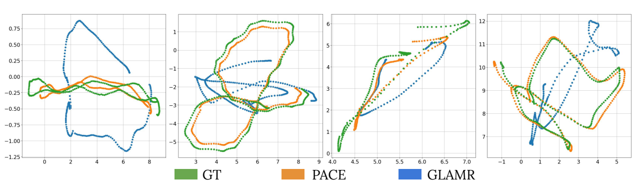


Figure 5. Comparison of camera motion estimation on HCM dataset. PACE estimates more accurate camera motions compared to GLAMR.

which shows that it is essential to use the background scene features in our unified optimization framework. During our experiments, we also evaluated the case when all variables are optimized from the beginning without stagewise optimization. We found that the optimization does not converge at all in this case.

5. Conclusion

We presented PACE, a novel approach for accurate global human and camera motion estimation from dynamic cameras. Our approach leverages the complementary benefits of human motion priors and SLAM methods and integrates them into a unified optimization framework that jointly optimizes human and camera motions. We also introduced a new synthetic dataset called HCM for benchmarking global human and camera motion estimation. We demonstrated that our approach achieves superior performance as compared to the state-of-the-art methods in accurately recovering both human and camera motion.

Although our method can refine camera trajectories obtained from SLAM, it may not be effective in scenarios where SLAM methods fail catastrophically. We believe that the integration of physics-based constraints to prevent camera errors from overriding human motion priors would be an interesting future direction. Another limitation of our method is the assumption of a planar ground caused by the lack of scene annotation in the AMASS dataset. Also, while our proposed optimization is efficient, it is not real-time and requires batch processing to exploit future and past tempo-

Methods	W-MPJPE ↓	WA-MPJPE ↓	PA-MPJPE ↓	ACCEL ↓	Runtime (per 1000 imgs)
GLAMR [107]	416.1	239.0	114.3	173.5	7min
SLAHMR [105]	**141.1**	101.2	79.13	25.78	400min
PACE (Ours)	147.9	**101.0**	**66.5**	**6.7**	8min

Table 4. State-of-the-art results on the EgoBody dataset.

ral information. Jointly solving camera and human motion in real-time and online fashion is a significant challenge.

A. Appendix

In this appendix, we provide results on an additional dataset, EgoBody [120], and also provide additional implementation details.

A.1. Experiments on EgoBody dataset

EgoBody [120] is a large-scale dataset capturing ground-truth 3D human motions during social interactions in 3D scenes. EgoBody 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 coordinate frame. We follow [105] and use the validation split of the dataset for evaluation. We use DROID-SLAM with the ground-truth camera intrinsics provided by the dataset.

Table 4 compares PACE with the state-of-the-art methods GLAMR [107] and SLAHMR [105]. As the results indicate, PACE significantly outperforms GLAMR while achieving performance on par with SLAHMR in terms of accuracy. However, PACE offers a significant computational advantage over SLAHMR, being up to 50 times faster for a sequence with 1000 frames. Note that the runtime of SLAHMR grows linearly with the sequence length, whereas our runtime increases sub-linearly. This improvement in efficiency demonstrates the potential of PACE as a practical and effective solution for human and camera motion estimation from videos.

A.2. Global optimization implementation details

We empirically chose the weights of all error terms involved in the optimization, as summarized in Table 5.

A.3. HCM dataset generation

To create our HCM (Human and Camera Motion) dataset we used the characters from the RenderPeople [1] dataset with 3D scenes from the Unreal Engine Marketplace [2]. We manually labeled the navigable areas in each 3D scene *i.e.*, sufficiently large, unobstructed flat areas within the scene. To generate a sequence, we randomly selected a 3D scene and a navigable area within it. We also randomly chose the number of people to be animated in the scene, ranging from 1 to 8 individuals. For each person, we selected a motion sequence from the validation set of the AMASS [68] dataset. To ensure that each person's motion sequence was optimized for the scene, we iteratively added one person at a time. We optimized their global translation to ensure that they remained within the bounds of the navigable area and did not intersect with existing people in the scene. We also check the terrain height of the navigable area and adjusted each character's root translation accordingly to ensure they were at the correct height relative to the terrain. Finally, we rendered the animated 3D scene into a video sequence using a moving camera. To generate camera trajectories, we designed heuristics to replicate typical camera movements observed in everyday videos and professional movies. More specifically, we used dolly zoom, random arc motion towards a person, camera motions from the MannequinChallenge dataset [59], cameras tracking a specific person, etc. This approach allowed us to generate a diverse set of sequences with varying numbers of people and diverse body and camera motions. In total, we generated 25 video sequences for evaluation. Some examples can be seen in the project page. We believe our HCM dataset will be extremely useful for evaluating human and camera motion estimation methods and furthering research in this direction.

References

[1] Render People, 2020. https://hdrihaven.com/. 6, 9

[2] Unreal Engine Marketplace, 2022. https://www.unrealengine.com/marketplace/en-US/content-cat/assets/environments/. 6, 9

[3] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *CVPR*, 2015. 2

[4] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *ICCV*, 2019. 3

[5] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *CVPR Workshops*, 2018. 3

[6] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, 2014. 3

[7] 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

[8] Tianshu Zhang Buzhen Huang, Yuan Shu and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *3DV*, 2021. 3

[9] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, pages 387–404. Springer, 2020. 3

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee.

Stages	Opt. Variables	Error Functions	Learning rate (lr) & Weights
Stage-1	s, h_0, R_0, β	$E_{2D} + E_\beta$	$\text{lr} = 0.01, \lambda_{2D} = 0.001, \lambda_\beta = 1.0$
Stage-2	$s, h_0, R_0, \beta, \mathbf{z}_\Phi$	$E_{\text{body}} + E_{\text{scene}}$	$\text{lr} = 0.01, \lambda_{\text{2D}} = 0.001, \lambda_\beta = 1, \lambda_{\text{contact}} = 100, \lambda_{\text{height}} = 10, \lambda_{\text{VAE}} = 0.1, \lambda_{\text{consist}} = 1, \lambda_{\text{smooth}} = 1$
Stage-3	$s, h_0, R_0, \beta, \mathbf{z}_\Phi, \mathbf{z}_\phi$	$E_{\text{body}} + E_{\text{scene}}$	$\text{lr} = 0.01, \lambda_{\text{2D}} = 0.001, \lambda_\beta = 1, \lambda_{\text{contact}} = 100, \lambda_{\text{height}} = 10, \lambda_{\text{VAE}} = 0.1, \lambda_{\text{consist}} = 1, \lambda_{\text{smooth}} = 1, \lambda_{\text{pose}} = 1$
Stage-4	$\beta, \mathbf{z}_\Phi, \mathbf{z}_\phi, R_t, T_t$	$E_{\text{body}} + E_{\text{scene}} + E_{\text{camera}}$	$\text{lr} = 0.001, \lambda_{\text{2D}} = 0.001, \lambda_\beta = 1, \lambda_{\text{contact}} = 100, \lambda_{\text{height}} = 10, \lambda_{\text{VAE}} = 0.1, \lambda_{\text{consist}} = 1, \lambda_{\text{smooth}} = 1, \lambda_{\text{pose}} = 1, \lambda_{\text{PCL}} = 1e^{-4}$

Table 5. Optimization stages and weights.

Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 2

[11] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, 2021.

[12] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 2

[13] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. Gravity-aware monocular 3d human-object reconstruction. In *ICCV*, 2021. 2

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

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

[16] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. *TPAMI*, 2021. 3

[17] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *ECCV*, 2020. 3

[18] Zijian Dong, Jie Song, Xu Chen, Chen Guo, and Otmar Hilliges. Shape-aware multi-person pose estimation from multi-view images. In *ICCV*, 2021. 3

[19] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *CVPR*, June 2020. 3

[20] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, 2015. 3

[21] 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. 2

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

[23] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *CVPR*, 2019. 3

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

[25] 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. 3

[26] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 3

[27] Nils Hasler, Bodo Rosenhahn, Thorsten Thormahlen, Michael Wand, Jürgen Gall, and Hans-Peter Seidel. Markerless motion capture with unsynchronized moving cameras. In *CVPR*, 2009. 3

[28] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 3

[29] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, pages 2282–2292, 2019. 3

[30] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. In *NeurIPS*, 2022. 2, 3, 4, 5

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

[32] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *CVPR*, 2019. 3

[33] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *3DV*, 2021. 3

[34] 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, 3, 6

[35] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *CVPR*, 2023. 3

[36] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *CVPR*, 2020. 2

[37] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. KAMA: 3D keypoint aware body mesh articulation. In *3DV*, 2021. 2, 3

[38] Mariko Isogawa, Ye Yuan, Matthew O'Toole, and Kris M Kitani. Optical non-line-of-sight physics-based 3d human pose estimation. In *CVPR*, 2020. 2

[39] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal

graphs. In *CVPR*, 2016. 3

[40] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 2

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

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

[43] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 5

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

[45] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *3DV*, 2020. 3

[46] Tarasha Khurana, Achal Dave, and Deva Ramanan. Detecting invisible people. In *ICCV*, pages 3174–3184, 2021. 3

[47] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[48] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2, 3

[49] 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

[50] 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

[51] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019.

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

[53] Jogendra Nath Kundu, Mugalodi Rakesh, Varun Jampani, Rahul Mysore Venkatesh, and R. Venkatesh Babu1. Appearance consensus driven self-supervised human mesh recovery. In *ECCV*, 2020.

[54] 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

[55] 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, 3

[56] Jiaman Li, Ruben Villegas, Duygu Ceylan, Jimei Yang, Zhengfei Kuang, Hao Li, and Yajie Zhao. Task-generic hierarchical human motion prior using vaes. In *3DV*, 2021. 4

[57] Jiaman Li, Ruben Villegas, Duygu Ceylan, Jimei Yang, Zhengfei Kuang, Hao Li, and Yajie Zhao. Task-generic hierarchical human motion prior using vaes. In *3DV*, 2021. 5

[58] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021. 2, 3, 4, 7

[59] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 9

[60] 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. 3

[61] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363*, 2017. 3

[62] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2

[63] 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, 6

[64] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 34(6):248:1–248:16, 2015. 3

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

[66] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *NeurIPS*, 34, 2021. 3

[67] Diogo Luvizon, Marc Habermann, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Scene-aware 3d multi-human motion capture from a single camera. In *EuroGraphics 2023*, 2023. 3

[68] 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, 9

[69] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017. 3

[70] 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. 2

[71] 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. In *SIGGRAPH*, 2020.

[72] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. In *SIGGRAPH*, 2017. 2

[73] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

thesis. In *ECCV*, 2020. 4

[74] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, 2019. 2

[75] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 2

[76] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self contact and human pose. In *CVPR*, 2021.

[77] 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.

[78] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 2

[79] 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

[80] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 2

[81] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *BMVC*, 2018. 3

[82] Christian Payer, Thomas Neff, Horst Bischof, Martin Urschler, and Darko Stern. Simultaneous multi-person detection and single-person pose estimation with a single heatmap regression network. In *ICCV PoseTrack Workshop*, 2017. 2

[83] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, 2021. 3

[84] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *ICCV*, 2019. 3

[85] N. Dinesh Reddy, Laurent Guigues, Leonid Pischulini, Jayan Eledath, and Srinivasa Narasimhan. Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In *CVPR*, 2021. 2, 3

[86] 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, 4

[87] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net: Localization-classification-regression for human pose. In *CVPR*, 2017. 2

[88] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *ICCV*, 2019. 2

[89] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. In *SIGGRAPH*, 2020. 2

[90] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[91] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 2

[92] 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. 2

[93] 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

[94] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, , and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, 2019. 2

[95] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *NeurIPs*, 2021. 4, 7

[96] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *ICLR 2023*, 2022. 3

[97] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017. 3

[98] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 3, 6

[99] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *CVPR*, 2021. 2

[100] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body and hands in the wild. In *CVPR*, 2019. 2

[101] 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. 2

[102] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPs*, 2022. 5

[103] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *ICCV*, 2019. 2

[104] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *ECCV*, 2018. 3

[105] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. 2, 3, 7, 9

[106] Hongwei Yi, Chun-Hao P Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J Black. Human-aware object placement for visual environment reconstruction. In *CVPR*, 2022. 2, 3

[107] 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, 3, 7, 8, 9

[108] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. In *ICLR 2020*, 2019. 3

[109] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 2020.

[110] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. In *NeurIPS*, 2020. 3

[111] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. *arXiv preprint arXiv:2212.02500*, 2022. 3

[112] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *CVPR*, 2021. 2

[113] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *ECCV*, 2020. 2

[114] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes the importance of multiple scene constraints. In *CVPR*, 2018. 2

[115] 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. In *NeurIPS*, 2018. 2

[116] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *ICCV*, 2021. 2, 3

[117] 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

[118] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, 2021. 2

[119] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3

[120] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, 2022. 9

[121] 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. 3

[122] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 2

[123] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4D association graph for realtime multi-person motion capture using multiple video cameras. In *CVPR*, 2020. 3

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

[125] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022. 4

[126] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. SMAP: Single-shot multi-person absolute 3d pose estimation. In *ECCV*, 2020. 3

[127] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *ICCV*, 2021. 3

[128] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular realtime full body capture with inter-part correlations. In *CVPR*, 2021. 2

[129] Yi Zhou, Jingwan Lu, Connelly Barnes, Jimei Yang, Sitao Xiang, and Hao Li. Generative tweening: Long-term inbetweening of 3d human motions. *ArXiv*, abs/2005.08891, 2020. 5