Synergistic Global-space Camera and Human Reconstruction from Videos

Yizhou Zhao[1]*, Tuanfeng Yang Wang[2], Bhiksha Raj[1], Min Xu[1], Jimei Yang[2], Chun-Hao Paul Huang[2]
[1]Carnegie Mellon University [2]Adobe Research
{yizhouz,bhiksha,mxu1}@cs.cmu.edu, {yangtwan,jimyang,chunhaoh}@adobe.com

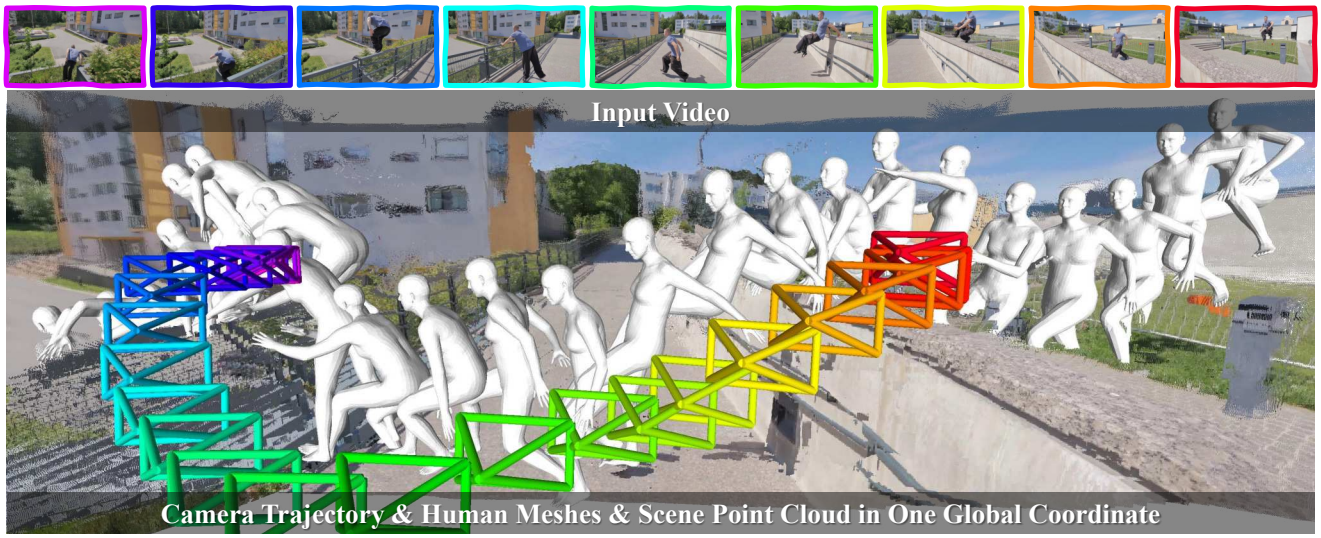

Figure 1. Our SynCHMR recovers metric-scale camera trajectories (color pyramids), human meshes, and scene point clouds in a global coordinate from casual videos by joining forces of Human Mesh Recovery (HMR) and Simultaneous Localization and Mapping (SLAM).

Abstract

Remarkable strides have been made in reconstructing static scenes or human bodies from monocular videos. Yet, the two problems have largely been approached independently, without much synergy. Most visual SLAM methods can only reconstruct camera trajectories and scene structures up to scale, while most HMR methods reconstruct human meshes in metric scale but fall short in reasoning with cameras and scenes. This work introduces Synergistic Camera and Human Reconstruction (SynCHMR) to marry the best of both worlds. Specifically, we design Human-aware Metric SLAM to reconstruct metric-scale camera poses and scene point clouds using camera-frame HMR as a strong prior, addressing depth, scale, and dynamic ambiguities. Conditioning on the dense scene recovered, we further learn a Scene-aware SMPL Denoiser to enhance world-frame HMR by incorporating spatio-temporal coherency and dynamic scene constraints. Together, they lead to consistent reconstructions of camera trajectories, human meshes, and dense scene point clouds in a common world frame.

1. Introduction

Physically plausible 3D human motion reconstruction from monocular videos is a long-standing problem in computer vision and graphics and has many applications in character animation, VFX, video games, sports, and healthcare. It requires estimating 3D humans across video frames in a common coordinate even with a moving camera. While human mesh recovery (HMR) has made significant progress recently [55], most existing methods typically estimate 3D humans in the camera coordinate by one frame at a time and fail to disambiguate camera motion. It calls for methods to jointly reconstruct 3D human and camera motion in a consistent global coordinate system from monocular videos. In other words, taking a video captured by a moving camera as input, the method should recover both temporally and spatially coherent movements of human bodies and cameras.

Intuitively, if the accurate camera motion is given, one can transform the bodies from individual camera frames to a common world frame by multiplying the inverse of camera extrinsic matrices. In practice, with humans moving in the scene, estimating the camera motion of a video is still an open challenge in monocular SLAM [1]. It not only

*Part of this work was done when interned at Adobe Research.

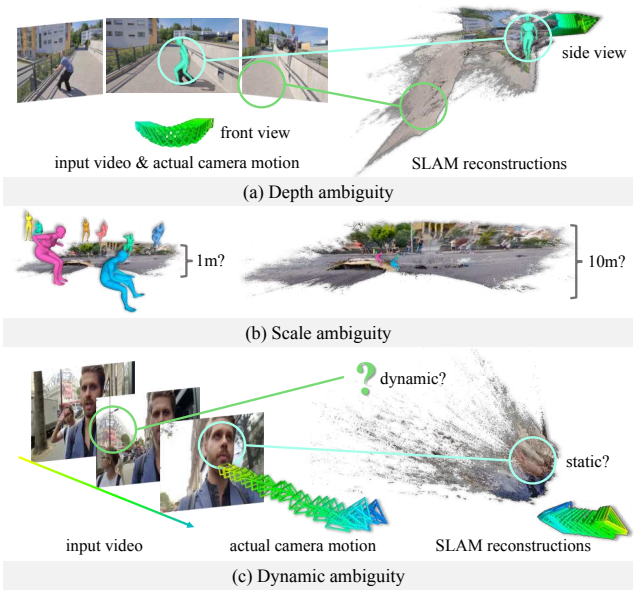

Figure 2. **Illustration of three types of ambiguities in visual SLAM.** We show SLAM reconstruction results from DROID-SLAM [54]. (a) Depth ambiguity occurs when there are only minor camera translations between different views. This can lead to geometric failures in reconstruction such as the folded back corridor in the side view. (b) Scale ambiguity is inherent in monocular SLAM systems and requires additional reference for disambiguation. (c) Dynamic ambiguity gets pronounced when moving foregrounds dominate frames. Over-reliance on foreground key points will result in incorrect camera trajectories.

falls short in capturing accurate depths on views with small camera translations but more crucially, only estimates scene structures and camera trajectories *up to scale*. The human motion also breaks the static key point assumption in the bundle adjustment. As a result, one needs additional reference to disambiguate the depth, the scale, and the dynamic as illustrated in Fig. 2.

To leverage SLAM results in HMR pipelines, current world-frame HMR methods often refine camera poses by integrating either partial camera parameters, such as a global scale of the translation [62], or full extrinsic matrices [15, 30, 46] in an optimization-based framework. However, their optimization-based nature leads to complex multi-stage schemes, making the overall pipeline unnecessarily slow and easy to break.

In this work, we explore a fundamentally different way to marry the best of HMR and SLAM. A 2D object can first be lifted from the image plane to the camera frame and then transformed into a common 3D space. This two-step process coincides with the combination of camera-frame HMR, which brings imaged 2D humans to 3D camera frames, and SLAM, which estimates the camera-to-world transformation. Noticing these correspondences, we leverage camera-frame HMR as a strong prior to bridge from the image plane to the camera frame for disambiguating SLAM, and utilize SLAM reconstructions to constrain the transformation of human meshes from individual camera frames to a common global space. The overall pipeline thus results in a better synergy of the two, which we dub Synergistic Camera and Human Reconstruction (SynCHMR).

We design SynCHMR based on several insights. First, despite camera-frame HMR methods cannot reconstruct humans in a coherent global frame, the estimated body dimension and location still provide cues to disambiguate SLAM. Unlike SLAHMR [62] which applies SLAM out of the box and corrects the scale afterward, we endow the SLAM process with human meshes from camera-frame HMR to address ambiguities. To this end, we capitalize on estimated *absolute* depths to provide pseudo-RGB-D inputs for SLAM [54] and confine the bundle adjustment to static backgrounds. Since current depth estimation methods [2, 43] predicts either relative depth maps or depths with data biases, we propose to calibrate their outputs by aligning with estimated human bodies in the camera frame [9]. With these priors, SLAM knows the depth, scale, and dynamic information from HMR and consequently estimates less ambiguous scene structures and camera poses.

Next, we place human meshes in the common coordinate recovered by SLAM. The gap between human tracks transformed from camera frames and their real plausible world-frame motions stems from two sources of error: noise induced by camera-frame HMR and by SLAM. Motion prior models [14, 44, 68] can be used for denoising purposes as they contribute to the temporal coherence of world-frame human tracks. However, their exclusive focus on human modeling either leaves them agnostic to the underlying scenes [14] or assumes the scene is a simple ground plane [44, 68]. Our intuition is that when placing a human, static elements of the scene, such as the ground, and dynamic components like moving objects are both possible to be in contact with the human, thereby providing clues for placing the body coherently and compatibly with the scene. We hence introduce a Scene-aware SMPL Denoiser that learns to denoise the transformed human tracks by considering both temporal consistencies of moving humans and implicit constraints from dynamic scenes. This global awareness makes it more flexible for in-the-wild videos.

Our contributions can be summed up as follows:

- We present a novel pipeline, SynCHMR, that takes a monocular video as input and reconstructs human motions, camera trajectory and *dense* scene point clouds all in one global coordinate, as shown in Fig. 1, whereas current world-frame HMR methods [30, 62, 65] can recover only an estimated or pre-defined ground plane.
- We propose a novel Human-aware Metric SLAM process to robustly calibrate estimated depth with estimated human meshes, resulting in metric-scale camera pose estimation and metric-scale scene reconstruction.

- We present Scene-aware SMPL Denoising that enforces spatiotemporal coherencies and applies dynamic scene constraints on world-frame human meshes. Notably, this is achieved without requiring extra annotations or heuristic designs to decide which part of a human should be interacting with the scene [49] and which region in the scene is most likely to be in contact with humans [38, 62].

2. Related Work

There is considerable prior arts of HMR. We briefly discuss how they adopt different camera models and refer the readers to [1] for a more comprehensive review.

HMR from a single image. State-of-the-art (SOTA) methods use parametric body models [21, 37, 41, 59] and estimate the parameters either by fitting to detected image features [3, 41, 58] or by regressing directly from pixels with deep neural networks [9, 11, 19, 22, 23, 28, 31, 32, 35, 45, 50, 60, 66, 67]. These approaches assume weak perspective/orthographic projection or pre-define the focal length as a large constant for all images. Kissos *et al*. [26] show that replacing focal length with a constant closer to ground truth alleviate the body tilting problem. SPEC [29] and Zolly [57] estimate focal length to account for perspective distortion. CLIFF [34] takes into account the location of humans in images to regress better poses in the camera coordinates.

Many of these camera-frame HMR methods assume zero camera rotation, which entangles body rotation and camera rotation. When applied on video data, they fail to reconstruct humans in a coherent global space since they operate in a per-frame manner and hence cannot reason about how the camera moves across frames.

HMR from videos aims to regress a series of body parameters from a temporal sequence. It opens up new problems such as whether the reconstructed bodies are in a common global coordinate or not. Some temporal methods consider a static camera [38, 44, 68], which makes the camera space a natural choice of the common coordinate. The challenge of coherent global space emerges when the camera moves. Early methods [5, 24, 27] show promising results on videos of dynamic cameras. Despite the reconstructed human meshes look great when overlaid on images, they do not share a common coordinate in 3D.

Recent HMR methods capitalize on human motion prior to constrain the global trajectories in the world space, which in turn implicitly disentangles human movement from camera movement. GLAMR [65] consider a data-driven prior models learned on large-scale MoCap database *e.g*. AMASS [39], while D&D [33] and Yu [64] consider physic-inspired prior. These world-frame HMR methods often struggle on noise in local poses caused by partial occlusions, which is very common in in-the-wild videos with close-up shots and crowd scenes. Kaufmann *et al*. [25] and BodySLAM++ [16] circumvent this problem by employing IMU sensors to provide more robust body estimates but require extra sensory devices. To fully disentangle human and camera motion, another line of work [15, 30, 36, 46] leverages state-of-the-art SLAM techniques, *e.g*. [47, 54, 70], to explicitly estimate camera motion from the input video and infer the body parameters in the world coordinate of SLAM. Closest to us is SLAHMR [62] which solves for a global scale to connect the pre-computed SLAM results and body trajectories. To carefully guide the optimization process, these methods tend to have complex, multi-stage optimization schemes, making the overall pipeline easy to break and unnecessarily slow.

Note that in stark contrast to the methods above, which either assume or estimate a simple ground plane as scene representation, SynCHMR reconstructs dense scenes from in-the-wild videos without pre-scanning with extra devices *a priori* like in [6, 7, 12, 13, 17, 61, 69]. We provide detailed comparisons with these world-frame HMR in Supp. Mat.

3. Method

Taking as input an RGB video $\{\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{T}$ with T frames and N people in the scene, we aim to recover human meshes $\{\mathbf{V}_{nt}^{\mathrm{w}} \in \mathbb{R}^{3\times 6890}\}_{n=1,t=1}^{N,T}$, dynamic scene point clouds $\{\mathbf{P}_t^{\mathrm{wm}} \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{T}$, and corresponding camera poses $\{\mathbf{G}_t^{\mathrm{m}} \in \mathrm{SE}(3)\}_{t=1}^{T}$ in a common world coordinate system. The superscripts $^{\mathrm{w}}$, $^{\mathrm{c}}$, and $^{\mathrm{m}}$ denote the world frame, the camera frame, and the metric scale, respectively. To this aim, we propose a two-phase alternative conditioning pipeline as depicted in Fig. 3. In the first phase, we calibrate camera motion by injecting a camera-frame human prior to SLAM. This resolves depth, scale, and dynamic ambiguities, yielding metric-scale camera poses and dynamic point clouds. Subsequently, in the second phase, we transform the camera-frame human tracks into the world frame and utilize the dynamic point clouds obtained in the first phase for conditional denoising.

3.1. Preliminaries

3.1.1 SLAM

Given a monocular RGB video $\{\mathbf{I}_t\}_{t=1}^{T}$, DROID-SLAM [54] solves a dense bundle adjustment for a set of camera poses $\{\mathbf{G}_t \in \mathrm{SE}(3)\}_{t=1}^{T}$ and inverse depths $\{\mathbf{d}_t \in \mathbb{R}_{+}^{H\times W}\}_{t=1}^{T}$. To update these estimations, it first computes a dense correspondence field $\mathbf{p}_{ij} \in \mathbb{R}^{H\times W\times 2}$ based on reprojection for each pair of frames (i, j):

$$\mathbf{p}_{ij} = \Pi(\mathbf{G}_{ij} \circ \Pi^{-1}(\mathbf{p}_i, \frac{1}{\mathbf{d}_i})), \tag{1}$$

where $\mathbf{p}_i \in \mathbb{R}^{H\times W\times 2}$ is a grid of pixel coordinates in frame i, $\mathbf{G}_{ij} = \mathbf{G}_j \circ \mathbf{G}_i^{-1}$ is the relative pose, and Π and Π^{-1} are the camera projection and inverse projection functions.

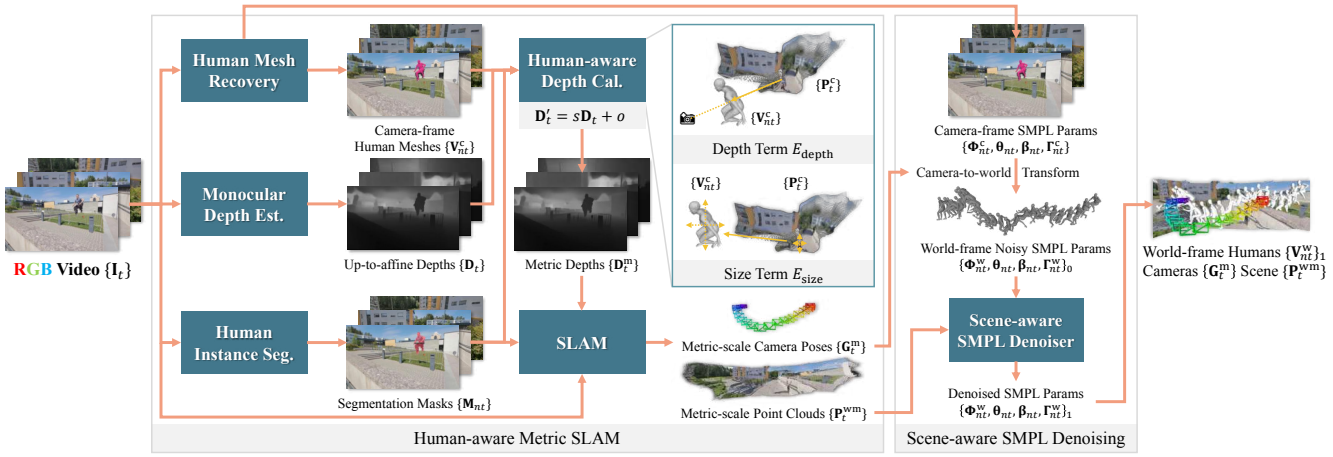

Figure 3. **The architecture of SynCHMR.** Our pipeline comprises two phases. The first phase, Human-aware Metric SLAM (Sec. 3.2), infers metric-scale camera poses and metric-scale point clouds by exploiting the camera-frame human prior. The second phase, Scene-aware SMPL Denoising (Sec. 3.3), involves the conditional denoising of world-frame noisy SMPL parameters. These parameters, initialized by transforming from the camera frame, get refined through conditioning on the dynamic point clouds obtained in the first phase. The whole pipeline thus reconstructs humans, scene point clouds, and cameras harmoniously in a common world frame.

Then with a learned neural network, the system predicts a revision flow field $\mathbf{r}_{ij} \in \mathbb{R}^{H\times W\times 2}$ and associated confidence map $\mathbf{w}_{ij} \in \mathbb{R}^{H\times W\times 2}_{+}$ to construct the cost function

$$E_\Sigma = \sum_{(i,j)} \left\| \mathbf{p}^*_{ij} - \Pi(\mathbf{G}'_{ij} \circ \Pi^{-1}(\mathbf{p}_i, \frac{1}{\mathbf{d}'_i})) \right\|^2_{\Sigma_{ij}}, \qquad (2)$$

where $\mathbf{p}^*_{ij} = \mathbf{r}_{ij} + \mathbf{p}_{ij}$ is the corrected correspondence, $\|\cdot\|_\Sigma$ is the Mahalanobis distance which weighs the error terms with $\Sigma_{ij} = \text{diag}\, \mathbf{w}_{ij}$, and $\mathbf{G}'$ and $\mathbf{d}'$ are updated poses and inverse depths. Upon this objective, DROID-SLAM considers an additional term that penalizes the squared distance between the measured and predicted depth if the input is with an extra sensor depth channel $\{\mathbf{D}_t\}_{t=1}^T$.

3.1.2 HMR

We employ 4DHumans [9] for reconstructing camera-frame human meshes from an in-the-wild video. Specifically, it performs per-frame human mesh recovery with an end-to-end transformer architecture and associates them to form human tracks. Each tracked human n in frame t is represented by SMPL [37] parameters as $\{\mathbf{\Phi}_{nt}, \boldsymbol{\theta}_{nt}, \boldsymbol{\beta}_{nt}, \mathbf{\Gamma}_{nt}\}$, including global orientation $\mathbf{\Phi}_{nt} \in \mathbb{R}^{3\times 3}$, body pose $\boldsymbol{\theta}_{nt} \in \mathbb{R}^{22\times 3\times 3}$, shape $\boldsymbol{\beta}_{nt} \in \mathbb{R}^{10}$, and root translation $\mathbf{\Gamma}_{nt} \in \mathbb{R}^3$. Then the parametric SMPL model can use these parameters to recover a human mesh with vertices $\mathbf{V}_{nt} \in \mathbb{R}^{3\times 6890}$ in metric scale: $\mathbf{V}_{nt} = \text{SMPL}(\mathbf{\Phi}_{nt}, \boldsymbol{\theta}_{nt}, \boldsymbol{\beta}_{nt}) + \mathbf{\Gamma}_{nt}$.

3.2. Human-aware Metric SLAM

3.2.1 Preprocessing

To start off, we estimate per-frame depth maps $\{\mathbf{D}_t\}$ with an off-the-shelf depth estimator, ZoeDepth [2] and predict per-frame human instance segmentation masks $\{\mathbf{M}_{nt}\}$ with an image instance segmentation network, Mask2Former [4]. We adapt ZoeDepth for video-consistent depth estimation by choosing a per-video metric head from the majority vote of per-frame routers, for which we dub ZoeDepth$^+$. While ZoeDepth claims to estimate metric depths, we observe a domain gap when inference on new datasets. Consequently, we only treat its output as up-to-affine depths that need to be further aligned with the metric scale. To aid our optimization with human awareness, we use camera-frame human meshes $\{\mathbf{V}^{\text{c}}_{nt}\}$ recovered by 4DHumans [9] to introduce a metric prior.

3.2.2 Calibrating Depth with Human Prior

We calibrate the per-frame depths with human meshes in Human-aware Depth Calibration. This involves optimizing two parameters, a world scale s and a world offset o, shared across all frames. During optimization, we linearly transform $\mathbf{D}_t$ to $\mathbf{D}'_t = s\mathbf{D}_t + o$ and unproject these depth maps to camera-frame point clouds $\{\mathbf{P}^{\text{c}}_t\}$ with $\mathbf{P}^{\text{c}}_t = \Pi^{-1}(\mathbf{p}_t, \mathbf{D}'_t)$. Our intuition is to align the human point cloud $\mathbf{P}^{\text{c}}_{nt} = \mathbf{M}_{nt} \odot \mathbf{P}^{\text{c}}_t$ with the camera-frame human mesh vertices $\mathbf{V}^{\text{c}}_{nt}$ in terms of absolute depth and size. To achieve pixel-wise alignment, we use a depth term to pull points on the human point cloud toward their corresponding human mesh vertices along the z-axis

$$E_{\text{depth}} = \frac{\sum_{n,t} \|\mathbf{S}_{nt} \odot [z(\mathbf{V}^{\text{c}}_{nt}) - z(\mathbf{P}^{\text{c}}_{nt})]\|^2_2}{\sum_{n,t} \|\mathbf{S}_{nt}\|_0}, \qquad (3)$$

where $\mathbf{S}_{nt} = \rho(\mathbf{V}^{\text{c}}_{nt}) \cap \mathbf{M}_{nt}$ is the intersection of the rasterized human mesh mask $\rho(\mathbf{V}^{\text{c}}_{nt})$ and the instance segmentation mask $\mathbf{M}_{nt}$, $z(\cdot)$ is the rasterized depth, and $\|\cdot\|_0$ is the

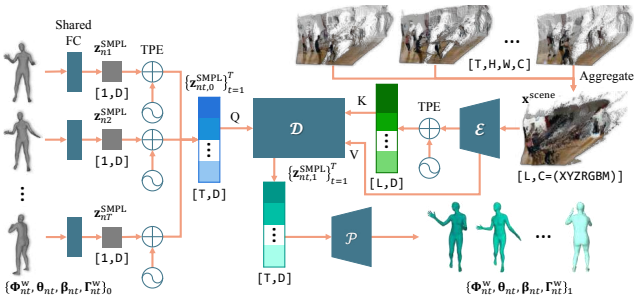

Figure 4. **The architecture of Scene-aware SMPL Denoiser.** World-frame noisy SMPL parameters $\{\mathbf{\Phi}^{\text{w}}_{nt}, \boldsymbol{\theta}_{nt}, \boldsymbol{\beta}_{nt}, \mathbf{\Gamma}^{\text{w}}_{nt}\}_0$ are first projected by a linear layer and summed with temporal positional embeddings (TPE) to get initial latent humans $\{\mathbf{z}^{\text{SMPL}}_{nt,0}\}$. Per-frame point clouds are aggregated to $\mathbf{x}_{\text{scene}}$ and encoded with the point encoder $\mathcal{E}$. Then we query the encoded scene $\mathcal{E}(\mathbf{x}^{\text{scene}})$ with latent humans $\{\mathbf{z}^{\text{SMPL}}_{nt,0}\}$ in the scene-conditioned denoiser $\mathcal{D}$ and feed the result $\{\mathbf{z}^{\text{SMPL}}_{nt,1}\}$ to prediction heads $\{\mathcal{P}_{\mathbf{\Phi}}, \mathcal{P}_{\boldsymbol{\theta}}, \mathcal{P}_{\boldsymbol{\beta}}, \mathcal{P}_{\mathbf{\Gamma}}\}$ to obtain denoised SMPL parameters $\{\mathbf{\Phi}^{\text{w}}_{nt}, \boldsymbol{\theta}_{nt}, \boldsymbol{\beta}_{nt}, \mathbf{\Gamma}^{\text{w}}_{nt}\}_1$.

0-norm indicating the number of non-zero pixels on a mask.

As the recovered human meshes can be noisy in depth but still have a stable body dimension, we also adopt a size term to leverage the relative position of mesh vertices

$$E_{\text{dx}} = \frac{\sum_{n,t} \|\Delta_x(\mathbf{V}^{\text{c}}_{nt}, \mathbf{S}_{nt}) - \Delta_x(\mathbf{P}^{\text{c}}_{nt}, \mathbf{S}_{nt})\|_2^2}{NT}. \quad (4)$$

We define E_{dy} similarly as E_{dx}, where

$$\Delta_*(\mathbf{X}, \mathbf{Y}) = (\max_* - \min_*) \left[\Pi^{-1}(\mathbf{Y} \odot \Pi(\mathbf{X}), z(\mathbf{X}))\right] \quad (5)$$

and $(\max_* - \min_*)$ denotes the difference between the maximum value and the minimum value on coordinate $*$.

Then we have the calibrated depths with optimization

$$(s^{\text{m}}, o^{\text{m}}) = \text{argmin}_{s,o} \left(E_{\text{depth}} + \lambda E_{\text{size}}\right), \quad (6)$$

$$\mathbf{D}^{\text{m}}_t = s^{\text{m}} \mathbf{D}_t + o^{\text{m}}, \quad (7)$$

where $E_{\text{size}} = E_{\text{dx}} + E_{\text{dy}}$, and λ is a hyperparameter to balance two energy terms with a default value of 1.

3.2.3 Disambiguating SLAM with Calibrated Depth

While DROID-SLAM [54] originally supports RGB-D input mode where the D channel stands for sensor depth, one cannot trivially access sensor depths from in-the-wild videos. Our insight is that an estimated absolute depth can be utilized as a depth prior, albeit noisy. So we combine the original RGB video and the calibrated depth as pseudo-RGB-D inputs $\{\mathbf{I}_t, \mathbf{D}^{\text{m}}_t\}$ to disambiguate depth and scale. Furthermore, we modify the cost function Eq. (2) to resolve the dynamic ambiguity by masking out dynamic foregrounds in confidence maps

$$\Sigma'_{ij} = \text{diag}\, \mathbf{w}'_{ij} = \text{diag}\left((1 - [\mathbf{M}_i, \mathbf{M}_j]) \odot \mathbf{w}_{ij}\right), \quad (8)$$

where $\mathbf{M}_i = \bigcup_n \mathbf{M}_{ni}$ and $\mathbf{M}_j = \bigcup_n \mathbf{M}_{nj}$ are the union of all human instance masks on their corresponding frame, and $[\cdot, \cdot]$ is the concatenation operation. As a result, we obtain metric-scale camera poses $\{\mathbf{G}^{\text{m}}_t\}$ and metric-scale point clouds $\{\mathbf{P}^{\text{wm}}_t\}$ by disambiguating SLAM with calibrated metric depths

$$\{\mathbf{G}^{\text{m}}_t, \mathbf{d}^{\text{m}}_t\} = \text{argmin}_{\{\mathbf{G}'_{ij}, \mathbf{d}'_{ij}\}} E_{\Sigma'}, \quad (9)$$

$$\mathbf{P}^{\text{wm}}_t = \mathbf{G}^{\text{m}}_t \circ \Pi^{-1}(\mathbf{p}_t, \frac{1}{\mathbf{d}^{\text{m}}_t})). \quad (10)$$

3.3. Scene-aware SMPL Denoising

3.3.1 Initializing Humans with Metric Cameras

To put humans properly in the scene recovered by SLAM, we initialize them by transforming estimated camera-frame SMPL parameters $\{\mathbf{\Phi}^{\text{c}}_{nt}, \boldsymbol{\theta}_{nt}, \boldsymbol{\beta}_{nt}, \mathbf{\Gamma}^{\text{c}}_{nt}\}$ to the world frame with camera-to-world transforms $\{\mathbf{G}^{\text{m}}_t = [\mathbf{R}_t | \mathbf{t}^{\text{m}}_t]\}$. Given the pelvis as the center of global orientation $\mathbf{\Phi}$, we have:

$$\mathbf{\Phi}^{\text{w}}_{nt} = \mathbf{R}_t \mathbf{\Phi}^{\text{c}}_{nt}, \quad \mathbf{\Gamma}^{\text{w}}_{nt} = \mathbf{R}_t(\mathbf{\Gamma}^{\text{c}}_{nt} + \mathbf{c}) + \mathbf{t}^{\text{m}}_t - \mathbf{c}, \quad (11)$$

where $\mathbf{c} = \mathbf{c}(\boldsymbol{\beta}_{nt})$ is the pelvis location in the shape blend body mesh. Note that we do not need to introduce an extra camera scale as SLAHMR [62] since the camera poses have already been in the metric scale. The root-relative poses $\boldsymbol{\theta}_{nt}$ and the shapes $\boldsymbol{\beta}_{nt}$ stay unchanged as in the camera frame. We denote the initialized and the denoised parameters with a suffix 0 and 1 respectively, *i.e.* $\{\mathbf{\Phi}^{\text{w}}_{nt}, \boldsymbol{\theta}_{nt}, \boldsymbol{\beta}_{nt}, \mathbf{\Gamma}^{\text{w}}_{nt}\}_{0,1}$.

3.3.2 Constraining Humans with Dynamic Scenes

Different from existing works [30, 36, 62, 64] that incorporate energy terms in optimization to apply explicit scene constraints, we propose to learn implicit scene constraints with a Scene-aware SMPL denoiser shown in Fig. 4. The noisy initial SMPL parameters $\{\mathbf{\Phi}^{\text{w}}_{nt}, \boldsymbol{\theta}_{nt}, \boldsymbol{\beta}_{nt}, \mathbf{\Gamma}^{\text{w}}_{nt}\}_0$ are first projected to a latent space, where it gets further updated by conditioning on implicit scene constraints

$$\mathbf{z}^{\text{SMPL}}_{nt,0} = \text{FC}\left([\mathbf{\Phi}^{\text{w}}_{nt,0}, \boldsymbol{\theta}_{nt,0}, \boldsymbol{\beta}_{nt,0}, \mathbf{\Gamma}^{\text{w}}_{nt,0}]\right) + \text{TPE}, \quad (12)$$

$$\{\mathbf{z}^{\text{SMPL}}_{nt,1}\}^T_{t=1} = \mathcal{D}\left(\{\mathbf{z}^{\text{SMPL}}_{nt,0}\}^T_{t=1}, \mathcal{E}(\mathbf{x}^{\text{scene}}) + \text{TPE}\right), \quad (13)$$

where FC is a shared linear layer, TPE is shared temporal positional embeddings, $\{\mathbf{z}^{\text{SMPL}}_{nt,*}\}^T_{t=1} \in \mathbb{R}^{T \times D}$ is the D-dimensional latent for human n, and $\mathbf{x}^{\text{scene}} \in \mathbb{R}^{L \times C}$ is the C-channel dynamic scene point clouds with a total number of points L. $\mathcal{E}$ and $\mathcal{D}$ refer to the scene encoder and the scene-conditioned denoiser, respectively. We set $C = 7$ which is the concatenation of point coordinates $\{\mathbf{P}^{\text{wm}}_t\}$, colors $\{\mathbf{I}_t\}$, and estimated human semantic segmentation masks $\{\mathbf{M}_t = \bigcup_n \mathbf{M}_{nt}\}$. Following [9], the updated latent

are decoded with different prediction head $\mathcal{P}_{(\cdot)}$ to regress the residual for each SMPL parameter:

$$\mathbf{\Phi}^{\text{w}}_{nt,1} = \mathcal{P}_{\mathbf{\Phi}}(\mathbf{z}^{\text{SMPL}}_{nt,1})\mathbf{\Phi}^{\text{w}}_{nt,0}, \tag{14}$$

$$\boldsymbol{\theta}_{nt,1} = \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{z}^{\text{SMPL}}_{nt,1})\boldsymbol{\theta}_{nt,0}, \tag{15}$$

$$\boldsymbol{\beta}_{nt,1} = \mathcal{P}_{\boldsymbol{\beta}}(\mathbf{z}^{\text{SMPL}}_{nt,1}) + \boldsymbol{\beta}_{nt,0}, \tag{16}$$

$$\mathbf{\Gamma}^{\text{w}}_{nt,1} = \mathcal{P}_{\mathbf{\Gamma}}(\mathbf{z}^{\text{SMPL}}_{nt,1}) + \mathbf{\Gamma}^{\text{w}}_{nt,0}. \tag{17}$$

We apply direct supervision on $\{\mathbf{\Phi}^{\text{w}}_{nt}, \boldsymbol{\theta}_{nt}, \boldsymbol{\beta}_{nt}, \mathbf{\Gamma}^{\text{w}}_{nt}\}_1$, which is common in the literature. Please see Supp. Mat. for the details of the full training objectives.

4. Experiments

4.1. Experimental Setting

Datasets. We assess the performance of SynCHMR primarily for global human motion estimation but also report the accuracy of estimated camera trajectories. Traditional video datasets in HMR literature are typically captured by static cameras, *e.g.* [13, 18, 20, 40, 63], hence not suitable for our purpose. Standard SLAM benchmarks such as [48, 51] do not meet our needs either as there is often no human moving in the scene. We consider the following datasets.

3DPW [56] is an in-the-wild dataset captured with iPhones. The ground truth bodies are not in coherent world frames so we use it to supervise root relative poses and for evaluation.

EgoBody [69] has ground-truth poses captured by multiple Kinects and egocentric-view sequences recorded by a head-mounted device, whose trajectories are further registered in the world space of Kinect array. We use it for training the SMPL denoiser in Sec. 3.3 and for evaluation (on both body and camera estimation). For HMR evaluation, unlike [30, 62] considering only the validation set, we additionally report results on its completely withheld test set.

EMDB [25] is a new dataset providing SMPL poses from IMU sensors and global camera trajectories. We include it for training the SMPL denoiser to enrich the diversity and use the camera trajectories to evaluate the quality of SLAM.

Evaluation Metrics. For HMR evaluation, we report common PA-MPJPE, which measures the quality of root-relative poses. For datasets that have ground-truth poses in a world coordinate, we follow [62] and consider WA-MPJPE and FA-MPJPE. The former measures the error after aligning the entire trajectories of the prediction and ground truth with Procrustes Alignment [10], while the latter aligns only with the first frame. We also report acceleration errors. For SLAM, we consider absolute trajectory error (ATE) for camera trajectory evaluation as well as the threshold accuracy (δ_n), the absolute relative error (REL), and the root mean squared error (RMSE) for scene depth evaluation [2].

Implementation Details. In Human-aware Depth Calibration, we use the L-BFGS algorithm with learning rate 1 to optimize for a maximum of 30 iterations. As for the Scene-aware SMPL Denoiser, we train it on the union of 3DPW-Train, EgoBody-Train, and EMDB for 100k steps with an AdamW optimizer, a batch size of 16, and a learning rate of 1e-5. For camera-frame SMPL ground truths like in 3DPW, we only incorporate body shapes $\boldsymbol{\beta}$ and poses $\boldsymbol{\theta}$ in training. We train the denoising process by randomly sampling a temporal window size T spanning 64 to 128 and inference with $T = 100$. The scene-conditioned denoiser $\mathcal{D}$ is parameterized with a 6-layer Transformer Decoder. For the scene encoder $\mathcal{E}$, we consider ViT and SPVCNN in Tab. 4 and report results for SPVCNN in Tabs. 1 and 2. Before inputting the world-frame noisy SMPL parameters to the denoiser, we first interpolate $\mathbf{\Phi}^{\text{w}}_{nt,0}$ and $\boldsymbol{\theta}_{nt,0}$ on $\text{SO}(3)$, $\boldsymbol{\beta}_{nt,0}$ on $\mathbb{R}^{10}$, and $\mathbf{\Gamma}^{\text{w}}_{nt,0}$ on $\mathbb{R}^3$ when there are missing observations.

Camera Model	Human Model	PA ↓
DROID-SLAM [54]	SLAHMR [62] w/ PHALP$^+$	55.9
DROID-SLAM [54]	SLAHMR [62] w/ 4DHumans [9]	57.4
Human-aware Metric SLAM (ours)	4DHumans [9]	52.9
Human-aware Metric SLAM (ours)	Scene-aware SMPL Denoiser (ours)	**52.4**

Table 1. **Comparison results on 3DPW-Test.** The row in gray is the full pipeline of SynCHMR. We abbreviate PA-MPJPE as PA, with the same below for FA-MPJPE (FA) and WA-MPJPE (WA).

4.2. Comparison Results

We first evaluate the estimated local poses with PA-MPJPE on 3DPW, which is common in the literature. In Tab. 1, we show that placing the bodies from 4DHumans already leads to lower error than SLAHMR. Passing them through the denoiser further reduces the error. We note that PA-MPJPE only measures local pose accuracy not the quality of global trajectories. Since 3DPW does not support any world metrics, Tab. 1 only aims to show that SynCHMR produces reasonable local poses on a common dataset.

Next, we assess the quality of global motion estimation, which is essentially a more challenging task. Tab. 2 shows the results on EgoBody. Note that current optimization-based methods [30, 62] report the error of the validation set. For fairness and completeness, we report results on both validation and test sets and run state-of-the-art methods on the test set when the code is available. In Tab. 2, we see that the proposed SynCHMR has the overall lowest PA-MPJPE, FA-MPJPE, and WA-MPJPE (gray rows). Comparing it with the row above (4DHumans) confirms the benefit of our scene-conditioned denoiser. For a fair comparison, we also initialize the global optimization of SLAHMR with 4DHumans, which is more accurate than PHALP$^+$ in SLAHMR, but we do not observe improvement. Notably, despite the concurrent work PACE [30] has a tightly integrated SLAM and body fitting objective, it still uses native DROID-SLAM to initialize the camera parameters like SLAHMR does. This is arguably sub-optimal as the initial-

Subset	Camera Model	Human Model	PA-MPJPE (mm) ↓	FA-MPJPE (mm) ↓	WA-MPJPE (mm) ↓	Acc Error (mm/frame^2) ↓	Runtime/100 imgs
Val	-	GLAMR [65]	114.3	416.1	239.0	173.5	4 min
	DROID-SLAM [54]	PACE [30]	66.5	147.9	101.0	**6.7**	1 min
	DROID-SLAM [54]	SLAHMR [62] w/ PHALP$^+$	79.1	141.1	101.2	25.8	40 min
	DROID-SLAM [54]	SLAHMR [62] w/ 4DHumans [9]	79.3	273.0	144.7	79.4	40 min
	Human-aware Metric SLAM (ours)	4DHumans [9]	73.0	164.4	106.7	127.0	5 min
	Human-aware Metric SLAM (ours)	Scene-aware SMPL Denoiser (ours)	**57.7**	**115.1**	**81.1**	64.8	5 min
Test	-	GLAMR [65]	112.8	351.4	216.3	105.9	4 min
	DROID-SLAM [54]	SLAHMR [62] w/ PHALP$^+$	63.1	163.9	99.4	**31.7**	40 min
	DROID-SLAM [54]	SLAHMR [62] w/ 4DHumans [9]	69.3	185.8	113.0	45.7	40 min
	Human-aware Metric SLAM (ours)	4DHumans [9]	75.4	160.0	108.1	138.8	5 min
	Human-aware Metric SLAM (ours)	Scene-aware SMPL Denoiser (ours)	**61.3**	**122.1**	**84.6**	69.4	5 min

Table 2. **Comparison results with state-of-the-art approaches on EgoBody.** The row in gray is the full pipeline of SynCHMR.

RGB	Depth	Mask	EgoBody ATE ↓	EgoBody δ_1 ↑	EgoBody REL ↓	EgoBody RMSE ↓	EMDB ATE ↓
✓	✗	✗	80.9	0.085	14.590	1617.361	400.3
✓	✗	Mask2Former [4]	81.6	0.063	8.530	1009.127	385.8
✓	ZoeDepth$^+$	✗	35.0	0.562	0.308	15.360	456.8
✓	ZoeDepth$^+$	Mask2Former [4]	28.6	0.564	0.307	10.852	389.6
✓	ZoeDepth$^+$ + Cal.	Mask2Former [4]	**26.4**	**0.797**	**0.274**	**10.452**	**107.0**

Table 3. **Ablation study for SLAM configurations in terms of optimized camera trajectories and scene depths.** ZoeDepth$^+$ denotes our video-adapted ZoeDepth [2].

Stage	Backbone	RGB	XYZ	Mask	PA ↓	FA ↓	WA ↓	Acc Error ↓
Init.	-	✗	✗	✗	73.7	120.8	93.1	127.1
Pred.	-	✗	✗	✗	63.3	98.8	77.2	75.2
Pred.	ViT [8]	✓	✗	✗	63.9	94.9	76.7	**43.3**
Pred.	ViT [8]	✓	✓	✗	64.5	97.3	77.7	45.6
Pred.	ViT [8]	✓	✗	✓	66.8	96.5	78.6	44.7
Pred.	ViT [8]	✓	✓	✓	69.3	100.9	82.0	46.4
Pred.	SPVCNN [53]	✗	✓	✗	62.9	95.1	76.0	72.6
Pred.	SPVCNN [53]	✓	✓	✗	**61.0**	93.4	74.3	67.7
Pred.	SPVCNN [53]	✗	✓	✓	62.0	93.9	75.3	69.9
Pred.	SPVCNN [53]	✓	✓	✓	61.3	**91.9**	**73.6**	64.8

Table 4. **Ablation study for different scene encoders and features regarding world-frame HMR.** Init. and Pred. refer to before and after SMPL denoising, respectively.

ization is not aware of body information, which can lead to errors that cannot be corrected in the global optimization stage. Consequently, it also has higher world-space errors. Optimization methods often employ a zero velocity term to smooth out human motion, which explains the lower acceleration error. However, we do not observe a big difference in jittery between our results. Please refer to Supp. Mat. for more details.

4.3. Ablation Study

We ablate the design choices in SynCHMR. In Tab. 3, we evaluate SLAM-optimized camera trajectories and scene depths with EgoBody and EMDB. We see that directly including un-calibrated monocular depths does not guarantee more accurate estimations (3rd vs. 1st and 4th vs. 2nd row). Precluding the dynamic foreground pixels with Mask2Former [4] generally improves performance. We empirically find that our depth calibration with human prior works the best when using it with foreground masking, which has the lowest error in both datasets. More SLAM evaluation and discussion can be found in Supp. Mat.

In Tab. 4, we verify the benefit of scene conditioning for the SMPL denoiser. We train it with EgoBody-train in different conditioning schemes and report the $T = 32$ results on EgoBody-val. First, placing the predicted bodies from 4DHuman in the global space directly with estimated camera extrinsics has the highest error (1st row). When conditioning on a constant zero tensor, the denoiser behaves like a motion prior and reduces the error (2nd row). To encode the appearance and geometry information of the scene, we consider ViT [8] or SPVCNN [53] as the encoder $\mathcal{E}$ and try varied combinations of appearance features (RGB), geometry features (XYZ) and aggregated subject masks (Mask). When using ViT to encode the scene, adding XYZ features or masks does not reduce the error. In contrast, when using SPVCNN, adding RGB information or conditioning on masks does improve performance. Overall, SPVCNN yields lower errors than ViT and enabling all conditioning leads to the lowest world-space error measure.

4.4. Qualitative Analysis and Discussion

In the first two rows of Fig. 5, we visualize the results of 3DPW and EgoBody in a global space. Despite occlusions, our SynCHMR estimates human meshes reliably and places them in a dense scene point cloud, whereas the scenes in GLAMR [65] and SLAHMR [62] consist of only a simple ground plane. Applying scene constraints with such an overly simplified scene can result in erroneous estimation, *e.g.*, incorrect human trajectories as shown in the top view of the 1st row, and the vertically shortened human bodies in the 4th row of (d). Note that since TRACE [52] is scene agnostic, the ground plane in (c) is only for visualization, not necessarily indicating scene penetration.

We also test on more in-the-wild DAVIS [42] videos containing human subjects. Since DAVIS provides no ground-truth human meshes nor camera trajectories, we show only the visual comparison. The 3rd row shows that we can handle multi-person cases as well as SLAHMR, while GLAMR often fails when multiple humans and dynamic cameras both occur. In a challenging scenario where the subject is taking selfies (the 4th row), both GLAMR and SLAHMR are confused by the foreground human dominating the frames and reconstruct an almost static global trajectory, failing to

(a) Input Video (b) GLAMR (c) TRACE (d) SLAHMR (e) SynCHMR (Ours)

Figure 5. **Qualitative comparison among world-frame HMR approaches.** We show (b) GLAMR [65] and (c) TRACE [52] results with their pre-defined ground planes, (d) SLAHMR [62] outputs with its estimated ground plane, and (e) our SynCHMR outputs with dense scenes. In the first row, we also demonstrate top-view human trajectories within circles. See supplementary for video results.

disentangle the camera and the human motions due to the dynamic ambiguity. TRACE fails to produce results due to severe frame truncation. In contrast, SynCHMR still successfully provides reasonable trajectories.

5. Limitation Discussion

As SynCHMR focuses on disentangling camera and human movements, we follow SLAHMR to approximate the focal length as $\frac{W+H}{2}$. When the subject has a shape that the body model cannot explain well, *e.g.*, children or obese people, calibrating depth with the estimated bodies is less ideal. As we develop and validate SynCHMR on real videos, its accuracy on composed or generated videos remains an open question. Finally, since SynCHMR handles dynamic scenes with moving subjects, it does not require an *a priori* scanned static scene. This opens up new challenges, such as incorporating dynamic point clouds as scene constraints.

6. Conclusion

We present SynCHMR, a method that reconstructs camera trajectories, human bodies, and dense scenes from in-the-wild videos all in one global coordinate. SynCHMR has two core innovations. First, it leverages monocular depth estimation and uses the dimension and location of human meshes to calibrate the range of depth. This allows SLAM to better resolve the inherent scale ambiguity problem as shown in the experiment. Second, we train a data-driven motion denoiser and condition it with the scene in the same global coordinate, which is the first such scene-conditioned motion prior. Combining the two, the full SynCHMR pipeline uses human bodies to improve SLAM, and the better estimated scene and camera trajectory, in turn, provide better constraints for feed-forward human motion denoising. It achieves SOTA results on common benchmarks compared with existing optimization-based approaches.

Acknowledgment

We appreciate constructive comments from Duygu Ceylan. This project was partially supported by the NIH under contracts R01GM134020 and P41GM103712, and by the NSF under contracts DBI-1949629, DBI-2238093, IIS-2007595, IIS-2211597, and MCB-2205148.

References

[1] Iman Abaspur Kazerouni, Luke Fitzgerald, Gerard Dooly, and Daniel Toal. A survey of state-of-the-art on visual SLAM. *Expert Systems with Applications*, 205:117734, 2022. 1, 3

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv*, 2023. 2, 4, 6, 7, 1

[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 *European Conference on Computer Vision (ECCV)*, 2016. 3

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 4, 7, 2

[5] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[6] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 6792–6802, 2022. 3

[7] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 682–692, 2023. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 7

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 4, 5, 6, 7, 1

[10] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 6

[11] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[12] 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 *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 3

[13] 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 (ICCV)*, pages 2282–2292, 2019. 3, 6, 1

[14] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. *Advances in Neural Information Processing Systems*, 35:4244–4256, 2022. 2

[15] Dorian F Henning, Tristan Laidlow, and Stefan Leutenegger. BodySLAM: joint camera localisation, mapping, and human motion tracking. In *European Conference on Computer Vision (ECCV)*, pages 656–673. Springer, 2022. 2, 3, 1

[16] Dorian F Henning, Christopher Choi, Simon Schaefer, and Stefan Leutenegger. BodySLAM++: Fast and tightly-coupled visual-inertial camera and human motion tracking. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 3781–3788. IEEE, 2023. 3

[17] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 3, 1

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2014. 6

[19] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[20] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *International Conference on Computer Vision (ICCV)*, 2015. 6

[21] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 3

[22] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*, pages 42–52, 2021. 3

[23] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

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

[25] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *International Conference on Computer Vision (ICCV)*, 2023. 3, 6

[26] Imry Kissos, Lior Fritz, Matan Goldman, Omer Meir, Eduard Oks, and Mark Kliger. Beyond weak perspective for

monocular 3d human pose estimation. In *European Conference on Computer Vision Workshops (ECCVw)*, pages 541–554, Cham, 2020. Springer International Publishing. 3

[27] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[28] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*. IEEE, 2021. 3

[29] 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 *International Conference on Computer Vision (ICCV)*, pages 11035–11045, 2021. 3

[30] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. PACE: Human and motion estimation from in-the-wild videos. In *International Conference on 3D Vision (3DV)*, 2024. 2, 3, 5, 6, 7, 1

[31] 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 *International Conference on Computer Vision (ICCV)*, 2019. 3

[32] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021. 3

[33] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&D: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 1, 2

[34] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, pages 590–606, 2022. 3

[35] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Computer Vision and Pattern Recognition (CVPR)*, pages 21159–21168, 2023. 3

[36] 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 *International Conference on 3D Vision (3DV)*, pages 930–939. IEEE, 2021. 3, 5, 1, 2

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34 (6):248:1–248:16, 2015. 3, 4

[38] Diogo Luvizon, Marc Habermann, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Scene-Aware 3D Multi-Human Motion Capture from a Single Camera. *Computer Graphics Forum (CGF)*, 42(2):371–383, 2023. 3

[39] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. 3

[40] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *International Conference on 3D Vision (3DV)*, 2018. 6

[41] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[42] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 7, 2

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(3):1623–1637, 2020. 2

[44] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3D human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 11468–11479, 2021. 2, 3

[45] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: Fast monocular 3d hand and body motion capture by regression and integration. In *International Conference on Computer Vision Workshops (ICCVw)*, 2021. 3

[46] Nitin Saini, Chun-Hao P Huang, Michael J Black, and Aamir Ahmad. SmartMocap: Joint estimation of human and camera motion using uncalibrated rgb cameras. *IEEE Robotics and Automation Letters*, 2023. 2, 3, 1

[47] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 1, 2

[48] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Computer Vision and Pattern Recognition (CVPR)*, pages 134–144, 2019. 6

[49] Zehong Shen, Zhi Cen, Sida Peng, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Learning human mesh recovery in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17038–17047, 2023. 3

[50] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[51] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *International Conference on Intelligent Robots and Systems (IROS)*, 2012. 6, 1, 3

[52] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D temporal regression of avatars with dynamic cameras in 3d environments. In *Computer Vision and Pat-*

tern Recognition (CVPR), pages 8856–8866, 2023. 7, 8, 1, 2

[53] Haotian* Tang, Zhijian* Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, 2020. 7

[54] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3, 5, 6, 7, 1

[55] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923*, 2022. 1

[56] 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 *European Conference on Computer Vision (ECCV)*, pages 614–631, 2018. 6

[57] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[58] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[59] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020. 3

[60] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3d pose and shape estimation by dense render-and-compare. In *International Conference on Computer Vision (ICCV)*, 2019. 3

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

[62] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 21222–21232, 2023. 2, 3, 5, 6, 7, 8, 1

[63] Jae Shin Yoon, Zhixuan Yu, Jaesik Park, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions and benchmark challenge. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(1):623–640, 2021. 6

[64] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *Transactions on Graphics (TOG)*, 40(6), 2021. 3, 5, 1, 2

[65] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11028–11039, 2022. 2, 3, 7, 8, 1

[66] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[67] 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 *International Conference on Computer Vision (ICCV)*, 2021. 3

[68] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4D human body capture in 3D scenes. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[69] 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 *European Conference on Computer Vision (ECCV)*, 2022. 3, 6

[70] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision (ECCV)*, pages 20–37. Springer, 2022. 3