This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# HumanBA: Human-Aware Bundle Adjustment via Global Human-Camera Decoupling

Fengyuan Yang, Tanuj Sur, Tze Ho Elden Tse, Angela Yao
National University of Singapore
{fyang, tanujsur, eldentse, ayao}@comp.nus.edu.sg

## Abstract

*Recovering global human and camera motion from monocular video is essential for world-coordinate human reconstruction but remains challenging due to entangled motions in image space. Traditional SLAM methods estimate monocular camera motion but fail in scenes dominated by foreground objects such as humans. A common workaround is to mask out dynamic objects, yet this approach becomes brittle when humans occupy most of the view or the background is too noisy, leading to unstable tracking and loss of constraints. This paper takes the opposite stance and reintegrates human motion as informative landmarks. We introduce* ***HumanBA****, a human-aware bundle adjustment framework that transforms dynamic humans into usable constraints via motion decoupling. HumanBA subtracts the human-induced component from observed joint trajectories, isolating a camera-induced (pseudo-static) component that can be safely incorporated into bundle adjustment alongside background features. To mitigate noise in global human estimates, HumanBA applies motion refinements and motion-aware reliability weighting. Across EMDB and SLOPER4D benchmarks, we show consistent improvements on camera pose estimation and reduce global human reconstruction error, demonstrating the benefits of treating humans as dynamic yet informative landmarks. Our code is available at* https://github.com/MartaYang/HumanBA.

## 1. Introduction

Recovering humans and cameras in world coordinates from monocular video requires disentangling the camera motion from the human motion [9, 30, 39, 44]. In human-centric videos, these motions are entangled in image space, which challenges classic Simultaneous Localization and Mapping (SLAM) pipelines [8, 34] that assume a largely static scene. As a result, world-coordinate human mesh recovery (HMR) methods typically choose between two imperfect strategies:

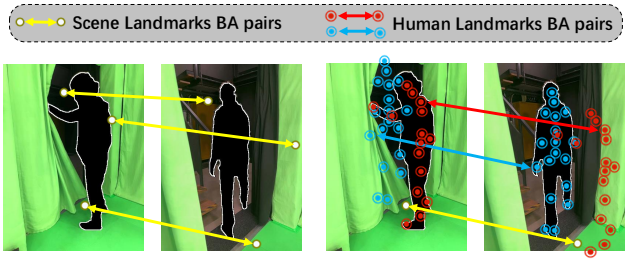



Figure 1. (a) Masked SLAM: Bundle adjustment (BA) optimizes only static landmarks (yellow) by excluding dynamic objects. (b) Proposed approach: Pseudo-static human landmarks (red and blue) are incorporated into BA by separating camera-induced and human-induced motion in observed trajectories.

(i) predict global human motion directly from local observations using learned local-to-global mappings [30, 33, 46], or (ii) run SLAM while masking dynamic humans to preserve static-scene assumptions [39]. The former can be inherently ambiguous, while the latter can remove valuable constraints and cause tracking dropouts when the foreground occupies most of the view or background scene is noisy (*e.g.*, in cases like Fig. 1a, masked SLAM will fail). This naturally raises the question: instead of treating the dynamic foreground human as outliers, can we leverage them as informative constraints to improve SLAM?

Recent dynamic-scene SLAM methods [3, 6, 45] support this perspective: dynamic foreground objects are not inherently problematic. For example, BA-Track [3] leverages dynamic 3D point tracker [1, 40] along with a motion decoupling mechanism to separate camera-induced motion from object motion. This allows bundle adjustment to jointly optimize camera and object trajectories on the isolated static motion components. While this improves robustness in dynamic scenes, BA-Track is object-agnostic and does not exploit human-specific priors, which are crucial for human-centric tasks such as global human and camera motion recovery.

Given human-centric video as input, our key observation is that, if human motion is properly modeled, it is not harmful to bundle adjustment. Furthermore, it can be highly ben-

eficial in noisy background scenarios (as shown in Fig. 1b). Bundle adjustment (BA) [37] jointly refines camera poses and 3D point positions by minimizing reprojection error across all frames and observations. Unlike local pose estimation, BA considers all frames and constraints together, which significantly improves accuracy and reduces drift. The real challenge lies in the mixture of camera-induced and human-induced motion in observed trajectories. When a human moves between frames, their apparent motion is a composite of both camera and human motion. Decoupling these components is highly non-trivial: human motion estimates are noisy and naive subtraction can amplify errors rather than reduce them. Making these dynamic elements usable as constraints requires careful modeling of motion priors and reliability.

To address this, we introduce HumanBA, a human-aware bundle adjustment framework that reintegrates human geometry into SLAM rather than suppressing it (as shown in Fig. 1b). Specifically, HumanBA leverages off-the-shelf human mesh recovery (HMR) models [36, 39, 44] to estimate the human-induced motion components (*e.g.*, the intra-frame movement from red to blue in Fig 1b). We then perform motion decoupling by subtracting these components from observed joint trajectories, isolating a pseudo-static, camera-induced motion signal (*e.g.*, the inter-frame movement of same color points in Fig. 1b). Since bundle adjustment is highly sensitive to the accuracy of reference constraints [8, 34], HumanBA assigns adaptive weights to human observations for each BA pair based on its motion stability over that pair's temporal span. These weighted joint observations are then incorporated into bundle adjustment alongside static background features, with unstable constraints downweighted to preserve robustness.

Our experiments show that HumanBA consistently reduces camera jumps and drifting error compared to existing methods. Moreover, by correcting global camera trajectories, HumanBA also reduces world-coordinate human motion errors, demonstrating a clear mutual benefit between human cues and camera optimization. We highlight our key contributions as follows:

- A human-aware bundle adjustment formulation that transforms dynamic human joints into usable constraints via motion decoupling.
- A confidence weighting mechanism that can adaptively score human observations based on motion stability.
- Improved camera and human reconstruction in world coordinates across different benchmarks.

## 2. Related Work

### 2.1. World Coordinate Human Mesh Recovery

Most HMR methods recover meshes in the camera coordinate frame from image inputs [4, 10, 18, 19, 23, 25, 27, 29, 38, 43]. Even video-based methods [5, 12, 20] largely remain camera-centric, adding temporal smoothness but lacking global grounding. Recent work has shifted toward world-coordinate HMR [9, 16, 22, 30, 33, 44, 46, 47], where the core challenge is disentangling camera motion from human motion. One class of methods learns local-to-global mappings that lift local pose and velocity cues to global human trajectories (*e.g.*, WHAM [30], TRACE [33], GLAMR [46], and DnD [16]). While effective in some cases, these predictions are inherently ambiguous—local observations (*e.g.*, walking on a treadmill) can correspond to vastly different global motions.

A second class of methods estimates camera motion explicitly, then composes it with local human predictions to obtain global trajectories (*e.g.*, SLAHMR [44], PACE [14], COIN [17], and SynCHMR [48]), typically using modern SLAM backends [24, 34, 35]. These approaches must still resolve monocular scale and handle foreground dynamics; joint optimization over scale, pose, and shape can be computationally heavy and prone to ambiguous minima. Recent systems differ mainly in how they obtain metric scale: TRAM [39] robustifies SLAM in human-centric videos and derives scale from background metric depth (*e.g.*, ZoeDepth [2]), whereas HAC [42] shows that foreground human depth—specifically contact joints from HMR—can serve as a reliable, optimization-free source of scale in the wild.

### 2.2. SLAM in Dynamic Scenes

Classical visual SLAM [8] assumes a mostly static world, meaning strong foreground motion violates multi-view consistency and degrades triangulation and bundle adjustment (BA). The standard remedy is to detect and suppress dynamic regions—mask moving objects, discard their features, or even inpaint the background—to preserve tracking and mapping (*e.g.*, DynaSLAM [1], StaticFusion for RGB-D [28]). While this stabilizes static maps, aggressive masking in human-centric videos often removes too many constraints, leading to brittle tracking and frequent failures.

Beyond outright masking, object-level dynamic SLAM segments and tracks multiple motions, reconstructing per-object geometry (*e.g.*, Co-Fusion [26], EM-Fusion [31]). These methods retain dynamic content but typically treat it as separate from camera optimization, so dynamic object motion rarely contributes to camera bundle adjustment. Recent dynamic-scene BA frameworks argue that dynamic elements are not inherently problematic—the real issue is the entanglement of camera-induced and object-induced motion. BA-Track [3] exemplifies this view by using a learning-based 3D point tracker to decompose observed motion and isolate the camera-induced component. This enables bundle adjustment to operate reliably even in highly dynamic scenes, yielding temporally coherent and

scale-consistent reconstructions. Our approach builds on this insight but differs in key ways: rather than relying on generic point tracking, we leverage structured human motion priors from off-the-shelf HMR models. We decouple human-induced motion from joint trajectories and use the residual pseudo-static component to support BA. Crucially, we introduce a motion stability-based confidence weighting mechanism that adaptively scores human constraints, allowing HumanBA to robustly incorporate human joints as landmarks in SLAM without requiring dense tracking or post-hoc depth refinement.

## 3. Preliminaries

### 3.1. Human Motion in Camera Coordinates

In video-based HMR, the 3D motion across $T$ frames is parameterized in the camera coordinate system by a sequence of SMPL variables $\mathcal{M}^c = \{\boldsymbol{\theta}_t, \boldsymbol{\beta}_t, \boldsymbol{\psi}_t, \boldsymbol{\tau}_t\}_{t=1}^T$, where SMPL [21] is a standard statistical model of human shape and pose. At each time step $t$, SMPL takes the body pose $\boldsymbol{\theta}_t \in \mathbb{R}^{23\times3}$, shape $\boldsymbol{\beta}_t \in \mathbb{R}^{10}$, root orientation $\boldsymbol{\psi}_t \in \mathbb{R}^3$, and root translation $\boldsymbol{\tau}_t \in \mathbb{R}^3$ and produces a set of 24 3D joints $\mathbf{J}_t = \mathcal{J}(\boldsymbol{\theta}_t, \boldsymbol{\beta}_t, \boldsymbol{\psi}_t, \boldsymbol{\tau}_t) \in \mathbb{R}^{24\times3}$, expressed in the camera coordinate system of frame $t$. Although SMPL names $\boldsymbol{\psi}_t$ and $\boldsymbol{\tau}_t$ the "global" orientation and translation, typical HMR methods [5, 10, 12, 13, 15, 41] estimate them with respect to the camera extrinsics $\mathcal{E} = \{\boldsymbol{R}_t, \boldsymbol{T}_t\}_{t=1}^T$, where $\boldsymbol{R}_t \in \mathbb{R}^{3\times3}$ and $\boldsymbol{T}_t \in \mathbb{R}^3$ denote the per-frame rotation and translation.

### 3.2. Human Motion in World Coordinates

Following [44, 46], one of our goals is to aggregate the per-frame local motions $\mathcal{M}^c$ across time into a single world frame[1] represented as $\mathcal{M}^w = \{\boldsymbol{\theta}_t, \boldsymbol{\beta}_t, \boldsymbol{\Phi}_t, \boldsymbol{\Gamma}_t\}_{t=1}^T$, which is decoupled from the per-frame camera extrinsics $\mathcal{E}_t$. Under standard perspective projection, the camera-frame root variables $\boldsymbol{\psi}_t, \boldsymbol{\tau}_t$ are obtained by applying $\mathcal{E}_t$ to the world-frame root orientation $\boldsymbol{\Phi}_t \in \mathbb{R}^3$ and translation $\boldsymbol{\Gamma}_t \in \mathbb{R}^3$:

$$\boldsymbol{\psi}_t = \boldsymbol{R}_t \boldsymbol{\Phi}_t; \quad \boldsymbol{\tau}_t = \boldsymbol{R}_t \boldsymbol{\Gamma}_t + \boldsymbol{T}_t. \tag{1}$$

To obtain global motion from local HMR predictions, we apply the inverse transform of the camera extrinsics to the local root:

$$\boldsymbol{\Phi}_t = \boldsymbol{R}_t^\top \boldsymbol{\psi}_t; \quad \boldsymbol{\Gamma}_t = \boldsymbol{R}_t^\top (\boldsymbol{\tau}_t - \boldsymbol{T}_t). \tag{2}$$

These equalities isolate and remove the camera motion when lifting trajectories into the world frame.

### 3.3. SLAM and Bundle Adjustment

We estimate camera extrinsics $\mathcal{E} = \{(\boldsymbol{R}_t, \boldsymbol{T}_t)\}_{t=1}^T$ and per-keyframe inverse depth (disparity) maps $\{d_i\}_{t=1}^T$ using visual SLAM. We use DROID-SLAM [34] as a representative

[1] We follow the convention that the first frame's camera extrinsics define the world coordinate system.

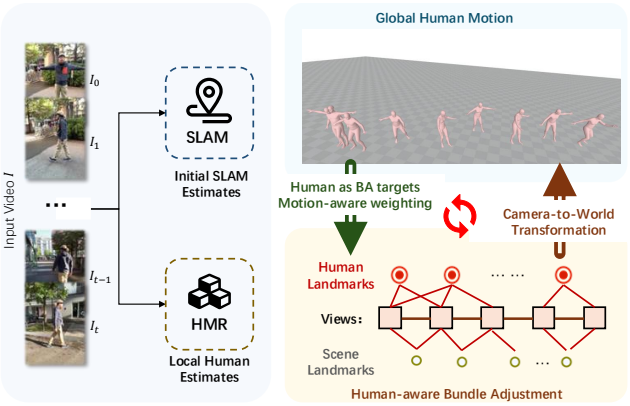



Figure 2. Framework overview. We first estimate human-induced motion from HMR models and SLAM, and isolate them from the observed trajectories to obtain human landmarks. We then assign adaptive weights based on motion stability and incorporate these weighted observations together with scene landmarks into BA. At each BA iteration, updated camera poses yield updated world-space human motion, enabling a mutually reinforcing human-aware bundle adjustment.

example, noting that many modern SLAM systems share similar bundle adjustment (BA) formulations. At the core of DROID-SLAM is a recurrent optimization over a frame graph, where each edge $(i, j) \in \mathbb{E}_{\text{pairs}}$ connects a pair of keyframes. For each pair, BA minimizes the reprojection error between a set of source pixels in frame $i$ and their predicted correspondences in frame $j$, under the current camera poses and depth estimates. The objective is:

$$\min_{\{\boldsymbol{R}_t, \boldsymbol{T}_t\}, \{d_i\}} \sum_{(i,j)\in\mathbb{E}_{\text{pairs}}} \left\| \mathbf{p}_{ij}^* - \Pi_c\big(\mathbf{G}_{ij} \circ \Pi_c^{-1}(\mathbf{p}_i, d_i)\big) \right\|_{\Sigma_{ij}}^2, \tag{3}$$

where $\Pi_c$ and $\Pi_c^{-1}$ denote projection and unprojection with intrinsics, $\mathbf{G}_{ij}(\boldsymbol{R}_i, \boldsymbol{T}_i, \boldsymbol{R}_j, \boldsymbol{T}_j)$ is the relative pose from $i$ to $j$, and $\Sigma_{ij} = \text{diag}(\mathbf{w}_{ij})$ stores per-pixel independent confidences predicted by the network. This formulation penalizes the discrepancy between the learned correspondence $\mathbf{p}_{ij}^*$ and the reprojection of 3D points from frame $i$ into frame $j$ under the current geometry. The pixel set $\mathbf{p}_i$ typically spans a dense grid at the internal feature resolution of the network. In dynamic scenes, foreground motion violates static multi-view assumptions and destabilizes BA. To mitigate this, masked variants of DROID-SLAM [39] suppress dynamic regions: human pixels are excluded from feature matching and assigned zero confidence ($\mathbf{w}_{ij} = 0$), effectively removing them from the optimization. Similar strategies are used in dynamic SLAM systems [1], which downweight or discard features inside moving-object masks. While this improves robustness, it also discards potentially informative constraints from dynamic foregrounds. Our method takes a different approach: instead of masking humans, we decouple their motion and reuse them as reliable landmarks for bundle adjustment.

## 4. Method

Our goal is to reintegrate masked humans back into bundle adjustment (BA) as pseudo-static landmarks that provide additional constraints to stabilize camera optimization. Fig. 2 outlines the pipeline. We first obtain initial SLAM and global human motion (Sec. 4.1). We then construct human-aware BA targets from the global motion and improve their accuracy with light smoothing/jump filtering (Sec. 4.2). Next, we assign adaptive weights to these human-aware BA targets based on motion stability over the temporal span (Sec. 4.3). Finally, we augment the standard BA objective by adding the human-aware term (Sec. 4.4).

### 4.1. Initial SLAM and Global Human Motion

We start from SLAM estimates of camera extrinsics $\mathcal{E}^{(0)} = \{(\boldsymbol{R}_t, \boldsymbol{T}_t)\}_{t=1}^{T}$ and a background point set $\mathcal{P} = \{\mathbf{x}'_n\}_{n=1}^{N}$, along with per-frame local HMR outputs $\mathcal{M}^c = \{(\boldsymbol{\theta}_t, \boldsymbol{\beta}_t, \boldsymbol{\psi}_t, \boldsymbol{\tau}_t)\}_{t=1}^{T}$. Using Eq. 2, we obtain an initial world-space human sequence $\mathbf{X}^w$, parameterized by $\{(\boldsymbol{\theta}_t, \boldsymbol{\beta}_t, \boldsymbol{\Phi}_t, \boldsymbol{\Gamma}_t)\}_{t=1}^{T}$. Metric scale can be calibrated based on the background scene [39] and the foreground human [42].

### 4.2. Pseudo-static Human as BA Targets

**General formulation of human-based BA targets.** As discussed in Sec. 3.3, the choice of reprojection targets $\mathbf{p}_i$ in Eq. 3 is critical. Masked DROID-SLAM excludes dynamic humans foregrounds by suppressing their features and setting their confidence weights to zero ($\mathbf{w}_{ij} = 0$). While this avoids multi-view violations, it also discards potentially informative constraints. We propose to reintegrate human observations into BA by treating selected human points as pseudo-static landmarks. These can be any reliably trackable 3D points on the human body—*e.g.*, SMPL joints [21], mesh vertices, or distinctive clothing features—provided their motion can be decoupled from camera motion. For each BA pair $(i, j) \in \mathbb{E}_{\text{pairs}}$, we define the human-based target in frame $j$ as the projection of landmarks from frame $i$, expressed in frame $j$ camera coordinates:

$$\mathbf{p}_{ij}^{\text{hum}*} = \Pi_c\big(\mathbf{X}_i^j\big), \tag{4}$$

where $\mathbf{X}_i^j$ denotes the human landmarks from frame $i$ transformed into frame $j$'s camera coordinate system. As illustrated in Fig. 3, we use SMPL joints [21] as $\mathbf{X}$, but the formulation is general.

**Motion decoupling for pseudo-static landmarks.** But how to get human at time $i$ under frame $j$'s camera coordinate $\mathbf{X}_i^j$ (red points on the right of the Fig. 3)? The answer is motion decoupling: the total observed motion of per-frame human in each camera coordinate $\mathbf{X}_i^i$ to $\mathbf{X}_j^j$ (purple arrow in Fig. 3) can be decomposed into a camera-induced part

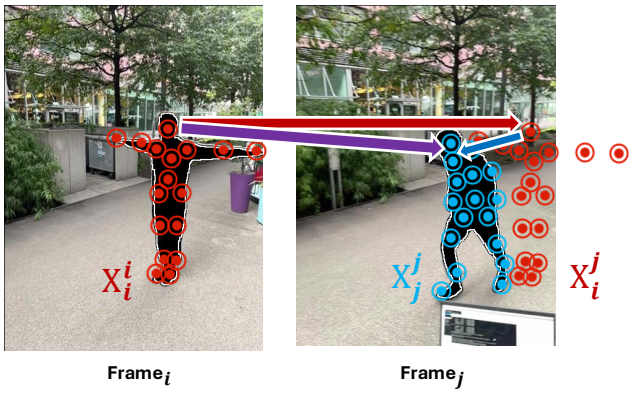



Figure 3. Example bundle adjustment pair $(i, j)$. The observed displacement from frame $i$ to $j$ (purple arrow) is caused by a camera-induced part (red arrow) and a human-induced part (blue arrow). By removing human-induced motion, we obtain a pseudo-static human $\mathbf{X}_i^j$ from time $i$ in frame $j$, which can serve as the human-aware BA target.

(red arrow) and a human-induced part (blue arrow). Given the current world human motion estimate $\mathbf{X}^w$ in Sec 4.1, the human-induced displacement in the world is

$$\Delta\mathbf{X}^w(i, j) = \mathbf{X}_j^w - \mathbf{X}_i^w, \tag{5}$$

whose expression in camera $j$ coordinate is:

$$\Delta\mathbf{X}^j(i, j) = \boldsymbol{R}_j\, \Delta\mathbf{X}^w(i, j). \tag{6}$$

Removing this human-induced motion (blue arrow in Fig. 3) from the observed human landmarks $\mathbf{X}_j^j$ in camera $j$ (blue points on the right of Fig. 3) yields the "backward-in-time" location of the human from frame $i$ in camera $j$ (red points on the right of Fig. 3):

$$\mathbf{X}_i^j = \mathbf{X}_j^j - \Delta\mathbf{X}^j(i, j). \tag{7}$$

Substituting Eq. 7 into Eq. 4 gives the final human-aware BA target:

$$\mathbf{p}_{ij}^{\text{hum}*} = \Pi_c\Big(\mathbf{X}_j^j - \boldsymbol{R}_j\big(\mathbf{X}_j^w - \mathbf{X}_i^w\big)\Big). \tag{8}$$

In short, we compute where the human landmarks would appear if only the camera moved by subtracting human-induced displacement from observed motion.

**Smoothing and filtering of global human motion.** To improve target accuracy, we refine the initial world trajectory $\mathbf{X}^w$ before forming human-aware targets. We detect abnormal steps by thresholding world-frame joint velocity/acceleration with a robust z-score, replace flagged increments by the local mean within a short temporal window, and then recompute the trajectory by cumulative integration, followed by light Gaussian smoothing. The refined world joints $\widehat{\mathbf{X}}^w$ replace $\mathbf{X}^w$ in the definitions above.

### 4.3. Adaptive human-aware BA Target Weight

Bundle adjustment is sensitive to the accuracy of reference landmarks [8, 34], so we assign an adaptive weight to each human-aware BA target according to motion stability over the temporal span $[i, j]$. We first define the interval-wise maximum discrete acceleration for each joint index $k \in [1, 24]$ in the world frame as:

$$a_{ij,k} = \max_{t \in [i,j]} \left\| \Delta^2 \widehat{\mathbf{X}}^w_{t,k} \right\|_2. \tag{9}$$

where $\Delta^2 \widehat{\mathbf{X}}^w_{t,k}$ is the acceleration magnitude of joint $k$. We then map $a_{ij,k}$ to a confidence with a bounded, monotone decreasing function:

$$w^{\text{hum}}_{ij,k} = \text{clamp}\Big(1 - \log\Big(\frac{\max(a_{ij,k}, \epsilon)}{\tau}\Big), 0, 1\Big), \tag{10}$$

where $\tau > 0$ controls the decay, $\epsilon$ is a small positive constant, and $\text{clamp}(x, 0, 1)$ saturates $x$ into $[0, 1]$. In the experiments section, we demonstrate empirically that this weighting behaves adaptively: joints near the root tend to have smaller $a_{ij,k}$ and receive larger weights than distal end-effectors (*e.g.*, hands), and long-span pairs $|j-i|$ automatically get smaller weights due to larger accumulated human-motion uncertainty. Both effects reduce the influence of unstable constraints while emphasizing reliable ones.

### 4.4. Human-aware Bundle Adjustment

We augment the background BA objective in Eq. 3 by adding human-landmark residuals on the same BA pair set $\mathbb{E}_{\text{pairs}}$. For each $(i, j) \in \mathbb{E}_{\text{pairs}}$ and each joint index $k \in [1, 24]$, we use the target $\mathbf{p}^{\text{hum}*}_{ij,k}$ defined in Eq. 8 and the geometry-induced projection obtained by transforming the joint from frame $i$ into frame $j$'s camera coordinates:

$$\ell^{\text{hum}}_{ij} = \sum_k w^{\text{hum}}_{ij,k} \left\| \mathbf{p}^{\text{hum}*}_{ij,k} - \Pi_c\big(\mathbf{G}_{ij} \circ \mathbf{X}^i_{i,k}\big) \right\|^2. \tag{11}$$

The total BA merges the background pixel residuals and our human-landmark residuals:

$$\min_{\{\boldsymbol{R}_t, \boldsymbol{T}_t\}, \{d_i\}} \sum_{(i,j) \in \mathbb{E}_{\text{pairs}}} \Big( \Big\| \mathbf{p}^*_{ij} - \Pi_c\big(\mathbf{G}_{ij} \circ \Pi_c^{-1}(\mathbf{p}_i, d_i)\big) \Big\|^2_{\Sigma_{ij}} + \ell^{\text{hum}}_{ij} \Big). \tag{12}$$

**Iterative update manner.** We follow the standard iterative BA schedule. After each update of the camera extrinsics $\{\boldsymbol{R}_t, \boldsymbol{T}_t\}$, we recompute the world-frame human trajectory using Eq. 2 (same lifting as in Sec. 4.1) and refresh the global human motion $\widehat{\mathbf{X}}^w$, the human-aware BA targets $\mathbf{p}^{\text{hum}*}_{ij,k}$, and the weights $w^{\text{hum}}_{ij,k}$ for the next BA iteration. This alternating refinement lets improved camera poses yield cleaner human anchors, which in turn provide more reliable BA targets in subsequent BA steps. Ours overall pseudo-code is shown in Algorithm 1.

**Algorithm 1** Human-aware Bundle Adjustment

**Require:** Initial SLAM extrinsics $\mathcal{E}^{(0)} = \{(\boldsymbol{R}_t, \boldsymbol{T}_t)\}_{t=1}^T$, inverse depths $\{d_i\}$, BA pairs $\mathbb{E}_{\text{pairs}}$, camera intrinsics $\Pi_c$, local HMR $\mathcal{M}^c$, iteration number $M$.
**Ensure:** Refined $\{\boldsymbol{R}_t, \boldsymbol{T}_t\}$ and global human motion $\widehat{\mathbf{X}}^w$.

```
X^w ← Initial global human from M^c and E^(0)        ▷ Eq. 2
X̂^w ← Motion-aware refinement of X^w                 ▷ Sec. 4.2
for m = 0 → M − 1 do
    p^hum*_{ij,k} ← Build human-aware BA targets for all (i, j) ∈ E_pairs   ▷ Eq. 8
    a_{ij,k} ← World-frame motion stability          ▷ Eq. 9
    w^hum_{ij,k} ← Motion-aware BA weights           ▷ Eq. 10
    ℓ^hum_{ij} ← Human-aware BA cost term            ▷ Eq. 11
    Update {R_t, T_t}, {d_i} by minimizing the total BA term   ▷ Eq. 12
    X^w ← Global human by new extrinsic              ▷ Eq. 2
    X̂^w ← Motion-aware refinement of X^w             ▷ Sec. 4.2
end for
```

## 5. Experiments

### 5.1. Implementation Details, Dataset and Metrics

**Datasets.** We evaluate the global human motion and camera motion on a subset of EMDB (EMDB2) [11] following previous global HMR works [30, 44]. EMDB2 contains 25 sequences captured by a dynamic camera and provides ground truth global motion for both the human and the camera. We also evaluate on SLOPER4D [7], an egocentric dataset captured in urban scenes, providing scene point clouds and global human/camera annotations. Since several sequences are very long, triggering the SLAM BA keyframe limit, we evaluate on a 3-sequence subset of the publicly released split (6 sequences in total) of this dataset, more details are provided in Sec. A of the Supplementary.

**Camera Motion Metrics.** Following SLAM convention [14, 32], we evaluate and report: (a) **ATE** (m), the Average Translation Error after rigidly aligning the camera trajectories; and (b) **ATE-S** (m), the Average Translation Error without scale alignment, providing a more accurate reflection of inaccuracies in the estimated scale of the scene.

**Human Motion Metrics.** We follow previous works [14, 30, 44] to evaluate global human motion in world coordinates with (a) **WA-MPJPE** (mm), the average Euclidean distance between the ground truth and the predicted joint positions (*i.e.* MPJPE) after aligning each segment for every 100 frames; (b) **W-MPJPE** (mm), the MPJPE error after only aligning the first two frames of each 100-frames segments with the ground truth; and (c) **RTE** (%), the human root translation error normalized by total sequence displacement after rigid alignment. For local mesh accuracy, we also report (d) **PA-MPJPE** (mm): MPJPE after per-frame procrustes alignment.

**Implementation Details**. We use DROID-SLAM [34] as our base SLAM system and run the human-aware BA for 19 iterations, the same as the original iteration number. Because DROID-SLAM operates on a $64 \times 48$ feature map, all 6890 SMPL vertices as BA targets is infeasible, so we select the 24 kinematic joints as landmarks. We adopt a z-score threshold $|z| > 3.5$ on world-frame joint velocity for global human motion refinement (Sec. 4.2). As for acceleration-based BA target weight in Eq. 10, we use a dataset-specific decay: $\tau$=1.0 on EMDB2 and $\tau$=2.0 on SLOPER4D [7].

## 5.2. Ablations

**Effectiveness of reintegrating Human in BA.** Tab. 1 compares three settings on EMDB2 [11]. *w/o Mask* (top block) performs worst in human-centric videos because strong foreground motion violates static-scene assumptions. *Masked SLAM* (middle block) stabilizes tracking by removing human pixels, but also discards useful constraints. *Reintegrating humans as BA targets* (bottom block) goes further: it reuses foreground humans as inliers via pseudo-static landmarks (Sec. 4.2), yielding additional gains over the masked baseline (last row). The improvement is larger after filtering sequences with pronounced inter-frame motion/acceleration spikes (Tab. 2, sequence details are provided in Sec. A of the Supplementary), indicating that HumanBA effectively corrects failure cases rather than merely smoothing already stable trajectories.

**Effect of human-motion refinement.** Human-motion refinement is crucial because bundle adjustment is sensitive to noisy references. In the bottom block of Tab. 1, using human joints as BA targets without motion-aware refinement (third row) yields smaller gains than with our refinement (last row). This indicates that denoising the global human trajectory successfully stabilizes optimization and improves accuracy.

**Which human joints as BA target?** Not all human parts are equally reliable as BA target constraints. Treating all joints equally (fourth row in Tab. 1) performs worse due to limb jitter and end-joints noise. Heuristically using only the root joint is more stable and can especially improve ATE-S. Going a step further, our adaptive target-weighting (Sec. 4.3, last row) achieves the best results, consistently improving camera metrics (ATE/ATE-S) and human metrics (W-MPJPE/WA-MPJPE).

**Analysis of the target-weighting mechanism.** We analyze learned weights across joints and temporal spans. As shown in Fig. 5a, the model assigns higher weights to proximal joints (*e.g.*, root, hip, and spine) and down-weights distal end-effectors (*e.g.*, hands), matching their relative stability. Fig. 5b shows that weights decrease as the pair span $|j-i|$ increases, reflecting larger accumulated uncertainty over longer intervals. Both trends align with the design goal of emphasizing reliable constraints and suppressing unstable ones.

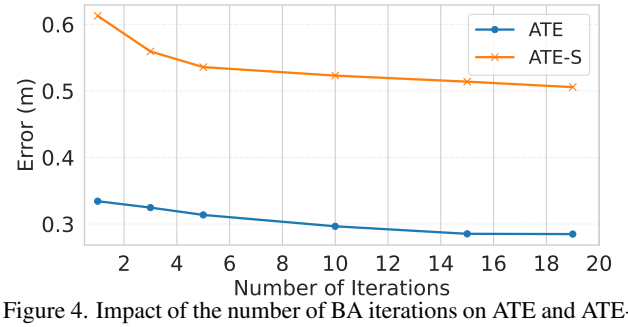
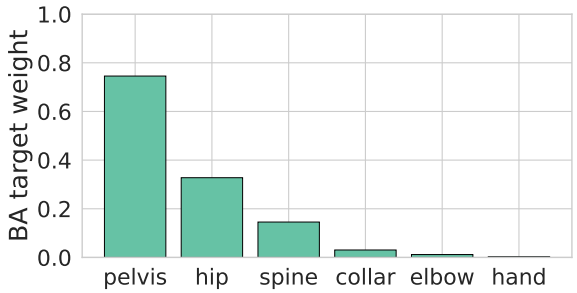
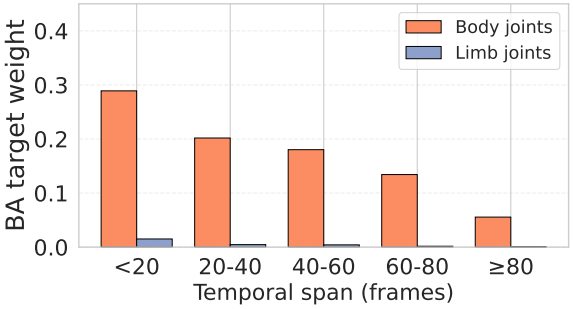



Figure 4. Impact of the number of BA iterations on ATE and ATE-S for EMDB2 [11] sequences with pronounced inter-frame motion spikes. Both metrics demonstrate a graceful decrease and quick convergence.




(a) Joint-wise HumanBA target weights averaged over the EMDB2. Proximal joints obtain higher weights; while distal joints (*e.g.*, wrists, hands) receive lower weights, consistent with their larger pose noise.




(b) HumanBA target weights vs. BA-pair temporal span $|j-i|$. The average weight decreases as the span grows, consistent with greater accumulated uncertainty over longer intervals.

Figure 5. Analysis of the target-weighting mechanism (Sec. 4.3).

**Iterative updates.** Fig. 4 plots the two camera evaluation metrics with respect to the number of bundle adjustment iterations on above mentioned sequences with abrupt inter-frame motion changes. The errors gracefully decrease with increasing iterations, converging to the final result after 8-10 iterations.

## 5.3. Comparison with the State-of-the-art

We benchmark our method against prior work and present two key findings. First, we show leveraging humans as BA targets improves SLAM robustness and accuracy, fixing trajectory jump failures in human-dominant scenes. Second, we demonstrate better camera estimates can feed back to

Table 1. Ablation studies of our methods on EMDB2 dataset [11]. The table consists of three ablation sections: (a) without Mask: no handling of dynamic humans; (b) Masked SLAM: ignore foreground humans; (c) Masked SLAM + Human-aware bundle adjustment: leveraging human joints as pseudo-static landmarks. Lower is better for all metrics. Best results are in **bold**.

| Dynamic Mask | Human Motion Refinement | Adaptive HumanBA Target Weight | Camera Metrics | | Human Metrics | |
|---|---|---|---|---|---|---|
| | | | ATE-S↓ | ATE↓ | W-MPJPE↓ | WA-MPJPE↓ |
| w/o Mask | – | – | 7.957 | 2.948 | 1071.49 | 421.67 |
| Masked DROID | – | – | 7.25 | 0.369 | 949.54 | 396.00 |
| Masked DROID + scale est. [39] | – | – | 0.708 | 0.369 | 230.97 | 79.80 |
| Masked DROID + HumanBA | ✗ | ✗ | 0.706 | 0.404 | 215.56 | 77.90 |
| Masked DROID + HumanBA | ✓ | ✗ (all joints) | 0.791 | 0.508 | 224.75 | 82.71 |
| Masked DROID + HumanBA | ✓ | ✗ (only root) | 0.718 | 0.398 | 211.86 | 74.81 |
| Masked DROID + HumanBA | ✓ | ✓ | **0.682** | **0.358** | **195.97** | **70.10** |

Table 2. Ablation studies of our methods on EMDB2 [11] sequences with pronounced inter-frame motion spikes.

| Ablation Study | Camera Metrics | | Human Metrics | |
|---|---|---|---|---|
| | ATE-S↓ | ATE↓ | W-MPJPE↓ | WA-MPJPE↓ |
| Masked DROID + scale est. [39] | 0.649 | 0.323 | 232.34 | 84.93 |
| Masked DROID + HumanBA (ours) | **0.525** | **0.285** | **193.69** | **74.49** |

Table 3. Camera pose error (ATE) on EMDB2 [11] by sequence length: short (<40 sec), medium (<60 sec), long (>60 sec), and Avg.(seq), the mean over all sequences. The top block lists ORB-SLAM2 [24] variants; the bottom block lists DROID-SLAM [34] variants. Gray rows denote numbers reported by original papers, while * denotes the errors we achieve from their model release.

| Methods | ATE↓ | | | |
|---|---|---|---|---|
| | Short | Medium | Long | Avg.(seq) |
| ORB-SLAM2 [24] | 0.08 | 0.29 | 2.08 | 1.05 |
| Masked ORB-SLAM2 [39] | 0.38 | 0.60 | 1.10 | 0.79 |
| DROID-SLAM [34] | 0.40 | 2.55 | 3.31 | 2.42 |
| Masked DROID [39] | 0.32 | 0.20 | 0.44 | 0.32 |
| Masked DROID* [39] | **0.24** | 0.25 | **0.47** | 0.34 |
| HumanBA (ours) | **0.24** | **0.21** | 0.48 | **0.32** |

Table 4. Comparison of scale-aware camera motion error (ATE-S) across sequence lengths on EMDB2 [11]. The evaluation protocol is the same as Tab. 3.

| Methods | ATE-S↓ | | | |
|---|---|---|---|---|
| | Short | Medium | Long | Avg.(seq) |
| TRAM [39] | 0.48 | 0.62 | 0.78 | 0.66 |
| TRAM* [39] | 0.51 | 0.62 | **0.81** | 0.67 |
| HumanBA (ours) | **0.37** | **0.45** | 0.90 | **0.61** |

stronger world-coordinate human reconstructions, confirming the mutual benefit highlighted in our method.

**Humans improve SLAM.** Tab. 3 shows that HumanBA yields the best overall ATE averages compared with prior SLAM methods (*e.g.*, ORB-SLAM2 [24]), DROID-SLAM [34], and their variants. Against masked DROID-SLAM, HumanBA outperforms on Short and Medium sequences and is on par on Long. For scale-aware evaluation in Tab. 4, HumanBA clearly surpasses TRAM [39] on Short/Medium and remains competitive on Long, leading to better overall averages. Tab. 5 evaluates our performance on a subset of SLOPER4D [7] as described in Sec. 5.1. In this more challenging dataset, HumanBA is substantially better than TRAM in ATE-S. These consistent ATE-S gains suggest that using humans as landmarks provides more reliable scale than background depth cues used by TRAM. Visualizations in Fig. 6 show that our method can mitigate the "masked SLAM" jumps and failure cases when the foreground human occupies large image regions.

Table 5. Comparison result of camera and human metrics on a subset of SLOPER4D [7] dataset. * denotes the errors we achieve from their model release.

| Methods | Camera Metrics | | Human Metrics | |
|---|---|---|---|---|
| | ATE-S↓ | ATE↓ | W-MPJPE↓ | WA-MPJPE↓ |
| TRAM* [39] | 4.05 | 1.89 | 736.90 | 131.42 |
| HumanBA (ours) | **3.10** | **1.86** | **674.45** | **122.67** |

**Better SLAM yields better global humans.** As shown in Tab. 6 and Tab. 5, HumanBA achieves lower world-coordinate human error (W-MPJPE and WA-MPJPE) and reduces root translation error than previous SOTA methods on two datasets. Noting that we do not have much improvement on PA-MPJPE (local pose metric) since local HMR is unchanged in our pipeline. However, the gains on world metrics indicate that improved camera motion by our human-aware BA substantially strengthens global human motion. This highlights a key insight: accurate global placement (camera translation and world alignment) often dominates global human accuracy—even when the local pose itself is unchanged. As shown in Fig. 7, world-space human visualization further shows HumanBA has smoother trajectories and fewer breaks compared to TRAM.

## 6. Conclusions and Outlook

We presented HumanBA, a human-aware bundle adjustment that reintegrates foreground humans into the SLAM

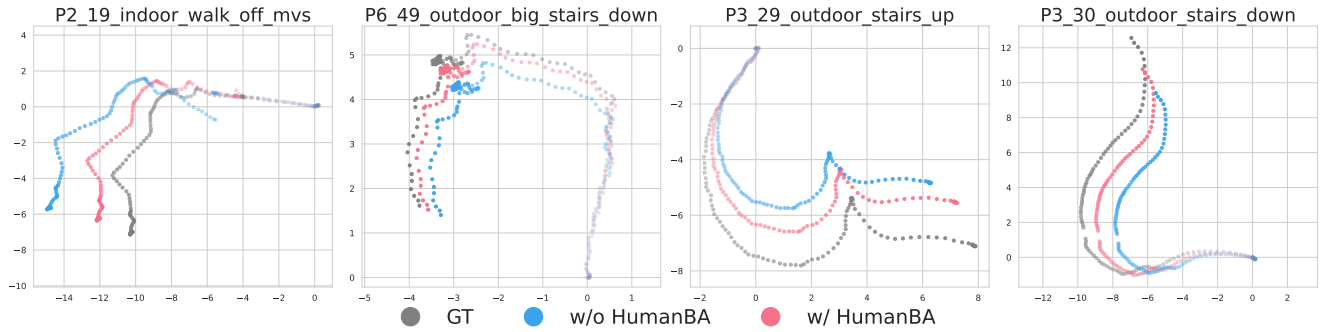



Figure 6. Camera trajectory before and after human-aware bundle adjustment. Our method shows less drifting and aligns more closely with ground truth than masked DROID-SLAM. Especially, HumanBA can mitigate SLAM failure like trajectory jump in the left-most case.

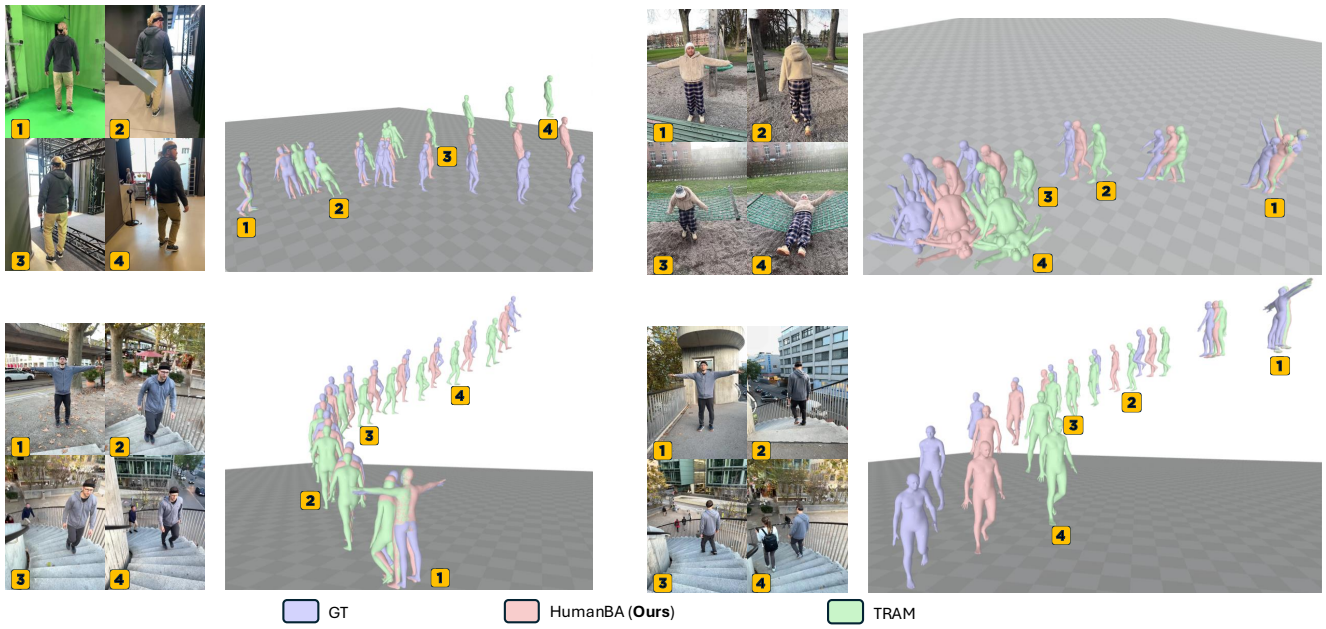



Figure 7. Qualitative examples of human motion in the world coordinate. We show that HumanBA can recover global human motions more accurately than the previous state-of-the-art method TRAM [39].

Table 6. SOTA comparison for global and local human motion on the EMDB2 [11]. * denotes the errors we achieve from TRAM's [39] model release.

| Models | EMDB2 | | | |
|---|---|---|---|---|
| | W-MPJPE↓ | WA-MPJPE↓ | PA-MPJPE↓ | RTE↓ |
| GLAMR [46] | 726.6 | 280.8 | 56.0 | 16.7 |
| TRACE [33] | 1702.3 | 529.0 | 58.0 | 18.9 |
| SLAHMR [44] | 776.1 | 326.9 | 61.5 | 10.2 |
| WHAM [30] | 354.8 | 135.6 | 41.9 | 6.0 |
| TRAM [39] | 222.4 | 76.4 | 38.1 | 1.4 |
| TRAM* [39] | 231.0 | 79.8 | **38.6** | 1.3 |
| HumanBA **(Ours)** | **196.0** | **70.1** | **38.6** | **1.2** |

optimization instead of masking them out. By decoupling observed motion into camera-induced and human-induced components, and constructing pseudo-static landmarks as BA targets on the existing factor graph, HumanBA stabilizes SLAM and, in turn, delivers stronger world-coordinate human estimation. Experiments show fewer trajectory jumps and drifting, and our adaptive acceleration-based weighting indeed emphasizes reliable joints and naturally down-weights long-span, uncertain pairs.

There are some settings on which our approach still struggles, giving clear indications as future work. First and foremost, our targets are derived from HMR; when there are strong occlusions, out-of-view joints, unusual apparel, or fast articulations, HMR itself struggles, which in turn may induce anchor drift. While our target weighting mechanism suppresses unstable joints, extreme cases may still have some impact. Secondly, our human-motion refinement currently adjusts only the root translation. If local HMR has catastrophic errors or is highly jittery, these errors can leak into the pseudo-static landmarks and weaken the BA benefits. Despite these limitations, our results confirm that HMR-modeled dynamic humans can serve as effective geometric constraints for SLAM when combined with carefully designed BA targets.

## Acknowledgement

This research is supported by A*STAR under its National Robotics Programme (NRP) (Award No. M23NBK0053).

## References


[1] Berta Bescos, José M Fácil, Javier Civera, and José Neira. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, 3(4):4076–4083, 2018. 1, 2, 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 preprint arXiv:2302.12288*, 2023. 2

[3] Weirong Chen, Ganlin Zhang, Felix Wimbauer, Rui Wang, Nikita Araslanov, Andrea Vedaldi, and Daniel Cremers. Back on track: Bundle adjustment for dynamic scene reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, 2025. 1, 2

[4] Hanbyel Cho, Yooshin Cho, Jaesung Ahn, and Junmo Kim. Implicit 3d human mesh recovery using consistency with pose and shape from unseen-view. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21148–21158, 2023. 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[6] Weichen Dai, Yu Zhang, Ping Li, Zheng Fang, and Sebastian Scherer. Rgb-d slam in dynamic environments using point correlations. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):373–389, 2020. 1

[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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 682–692, 2023. 5, 6, 7

[8] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE Robotics and Automation Letters*, 13(2):99–110, 2006. 1, 2, 5

[9] 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. 1, 2

[10] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[11] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2023. 5, 6, 7, 8

[12] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[13] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 3

[14] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and camera motion estimation from in-the-wild videos. *International Conference on 3D Vision (3DV)*, 2023. 2, 5

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

[16] 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)*, pages 479–496. Springer, 2022. 2

[17] Jiefeng Li, Ye Yuan, Davis Rempe, Haotian Zhang, Pavlo Molchanov, Cewu Lu, Jan Kautz, and Umar Iqbal. Coin: Control-inpainting diffusion prior for human and camera motion estimation. In *European Conference on Computer Vision (ECCV)*, pages 426–446. Springer, 2024. 2

[18] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *IEEE International Conference on Computer Vision (ICCV)*, pages 12939–12948, 2021. 2

[19] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021. 2

[20] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang Wang. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11006–11016, 2022. 2

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

[22] Diogo C Luvizon, Marc Habermann, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Scene-aware 3d multi-human motion capture from a single camera. In *Computer Graphics Forum (Eurographics)*, pages 371–383. Wiley Online Library, 2023. 2

[23] 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 *European Conference on Computer Vision (ECCV)*, pages 752–768. Springer, 2020. 2

[24] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2, 7

[25] 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 *IEEE International Conference on Computer Vision (ICCV)*, pages 5340–5348, 2019. 2

[26] Martin Rünz and Lourdes Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In

*2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4471–4478. IEEE, 2017. 2

[27] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 84–93, 2020. 2

[28] Raluca Scona, Mariano Jaimez, Yvan R Petillot, Maurice Fallon, and Daniel Cremers. Staticfusion: Background reconstruction for dense rgb-d slam in dynamic environments. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3849–3856. IEEE, 2018. 2

[29] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. *arXiv preprint arXiv:2009.10013*, 2020. 2

[30] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2070–2080, 2024. 1, 2, 5, 8

[31] Michael Strecke and Jorg Stuckler. Em-fusion: Dynamic object-level slam with probabilistic data association. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2019. 2

[32] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 573–580. IEEE, 2012. 5

[33] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8856–8866, 2023. 1, 2, 8

[34] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:16558–16569, 2021. 1, 2, 3, 5, 6, 7

[35] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[36] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923*, 2022. 2

[37] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999. 2

[38] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *European Conference on Computer Vision (ECCV)*, pages 20–38. Springer International Publishing, 2018. 2

[39] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision (ECCV)*, pages 467–487. Springer, 2024. 1, 2, 3, 4, 7, 8

[40] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024. 1

[41] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:38571–38584, 2022. 3

[42] Fengyuan Yang, Kerui Gu, Ha Linh Nguyen, Tze Ho Elden Tse, and Angela Yao. Humans as checkerboards: Calibrating camera motion scale for world-coordinate human mesh recovery. In *IEEE International Conference on Computer Vision (ICCV)*, 2025. 2, 4

[43] P. Yao, Z. Fang, F. Wu, Y. Feng, and J. Li. Densebody: Directly regressing dense 3d human pose and shape from a single color image. *arXiv preprint arXiv:1903.10153*, 2019. 2

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

[45] Chao Yu, Zuxin Liu, Xin-Jun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei. Ds-slam: A semantic visual slam towards dynamic environments. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 1168–1174. IEEE, 2018. 1

[46] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11038–11049, 2022. 1, 2, 3, 8

[47] Yuxiang Zhang, Hongwen Zhang, Liangxiao Hu, Jiajun Zhang, Hongwei Yi, Shengping Zhang, and Yebin Liu. Proxycap: Real-time monocular full-body capture in world space via human-centric proxy-to-motion learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1964, 2024. 2

[48] Yizhou Zhao, Tuanfeng Yang Wang, Bhiksha Raj, Min Xu, Jimei Yang, and Chun-Hao Paul Huang. Synergistic global-space camera and human reconstruction from videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1216–1226, 2024. 2