Reconstructing People, Places, and Cameras

Lea Müller* Hongsuk Choi* Anthony Zhang Brent Yi Jitendra Malik Angjoo Kanazawa
UC Berkeley
{mueller,hongsuk,anthony_zhang1234,brentyi,malik,kanazawa}@berkeley.edu
* equal contribution

Figure 1. **Humans and Structure from Motion (HSfM).** We propose a method for the joint reconstruction of humans, scene point clouds, and cameras from an uncalibrated, sparse set of images depicting people. By explicitly incorporating humans into the traditional Structure from Motion (SfM) framework through 2D human keypoint correspondences and leveraging robust initialization from an off-the-shelf model for scene and camera reconstruction, our approach demonstrates that integrating these three elements—people, scenes, and cameras—synergistically improves the reconstruction accuracy of each component. Unlike prior work in SfM and human pose estimation, our method reconstructs metric-scale scene point clouds and camera parameters, informed by human mesh predictions, while situating human meshes in coherent world coordinates consistent with the surrounding environment without any explicit contact constraints.

Abstract

We present "Humans and Structure from Motion" (HSfM), a method for jointly reconstructing multiple human meshes, scene point clouds, and camera parameters in a metric world coordinate system from a sparse set of uncalibrated multi-view images featuring people. Our approach combines data-driven scene reconstruction with the traditional Structure-from-Motion (SfM) framework to achieve more accurate scene reconstruction and camera estimation while simultaneously recovering human meshes. In contrast to existing scene reconstruction and SfM methods that lack metric scale information, our method estimates approximate metric scale by leveraging the human statistical model. Furthermore, our method reconstructs multiple human meshes within the same world coordinate system with the scene point cloud, effectively capturing spatial relationships among individuals and their positions in the environment. We initialize the reconstruction of humans, scenes, and cameras using robust foundational models and jointly optimize these elements. This joint optimization synergistically improves the accuracy of each component. We

*compare our method with existing methods on two challenging benchmarks, EgoHumans and EgoExo4D, demonstrating significant improvements in human localization accuracy within the world coordinate frame (reducing error from 3.59m to 1.04m in EgoHumans and from 3.01m to 0.50m in EgoExo4D). Notably, our results show that incorporating human data into the SfM pipeline improves camera pose estimation (*e.g.*, increasing RRA@15 by 20.3% on EgoHumans). Additionally, qualitative results show that our approach improves scene reconstruction quality. Our code is available at* muelea.github.io/hsfm.

1. Introduction

In recent years, combining deep learning with multi-view geometry has led to significant advances in two key areas: 3D human reconstruction [16, 23] and scene reconstruction [48, 53]. However, progress in these domains has largely evolved independently. Human reconstructions often lack anchoring within their surrounding scenes, while scene reconstructions typically exclude people and fail to recover metric scale. In this paper, we propose a unified framework that bridges these two elements.

We introduce Humans and Structure from Motion (HSfM), a new method that enables the joint reconstruction of multiple human meshes, scene point clouds, and camera parameters within the same metric world coordinate system as shown in Figure 1. From a sparse set of uncalibrated multi-view images featuring people, our approach combines data-driven scene reconstruction with the traditional Structure-from-Motion (SfM) framework to enhance the accuracy of scene and camera reconstruction while simultaneously estimating human meshes. The reconstruction process is initialized using robust foundational models for scene reconstruction [53] and human reconstruction [16] and further refined through joint optimization. This optimization incorporates a global alignment loss on scene pointmaps and bundle adjustment based on 2D human keypoint predictions [59], significantly enhancing the accuracy of the three components of world reconstruction—humans, scenes, and cameras. Our overall pipeline is depicted in Figure 2.

Unlike existing approaches to multi-view scene reconstruction [41, 42, 53, 60] and human pose estimation [8, 23, 55], HSfM recovers the metric scale of scene point clouds and camera poses while situating human meshes within a unified world coordinate system. The comprehensive output of HSfM facilitates the capture and evaluation of spatial relationships among individuals, ensuring consistency with the surrounding environment. Furthermore, unlike prior multi-view human pose estimation methods that depend on precise camera calibration [12, 20, 64], our approach operates with minimal constraints on the capture setup and does not require prior knowledge of the environment.

Our approach is founded on two key insights. The first insight is that deep learning-based human mesh estimation inherently contains metric scale information, as the predictions reflect the statistical human size present in the training datasets, thereby constraining the scale of the scene. The second insight is that robust 2D human keypoint predictions and 3D human mesh estimations provide precise correspondences and reliable initial 3D structures for bundle adjustment. Note that for the purpose of this work, we assume known re-identification of people across camera views.

We evaluate our approach on two challenging benchmarks, EgoHumans [24] and EgoExo4D [17], which feature individuals participating in a variety of indoor and outdoor activities across diverse environments. We assess the accuracy of human mesh reconstruction by comparing our method to other approaches that estimate human poses in a world coordinate frame [56]. Additionally, we compare camera pose accuracy against learning-based dense scene reconstruction methods, such as DUSt3R [53] and MASt3R [26]. Our approach demonstrates substantial improvements in camera pose estimation compared to existing methods while accurately positioning individuals within the scene. Specifically, it achieves approximately a 3.5-fold improvement in human metrics, reducing the human world location error from 3.51m to 1.04m on EgoHumans, and delivers camera metric improvements of approximately 2.5 times compared to the most relevant baseline [55]. These results underscore the effectiveness of our method, which leverages the joint reconstruction of multiple human meshes, scene point clouds, and cameras, supported by robust initialization for humans [16] and cameras [53]. We further validate our design through ablations which show the synergy between humans, scenes, and cameras. Our qualitative results highlight that the joint optimization of people (multiple humans), places (scenes), and cameras not only enhances human localization but also improves scene reconstruction and camera pose estimation.

In summary, we present Humans and Structure from Motion (HSfM), which provides a comprehensive representation of the world—encompassing people, places, and cameras—marking a step forward in understanding complex real world environments.

2. Related Work

Research in multi-person mesh reconstruction and Structure from Motion has seen substantial progress showing remarkable domain-specific results (see Tab. 1). Building on these foundational works, our approach unifies these areas.

Structure from Motion. Structure from Motion (SfM) [5, 15, 19] aims to reconstruct camera poses and 3D scene geometry from a set of images by establishing pixel correspondences across views. Traditional SfM Meth-

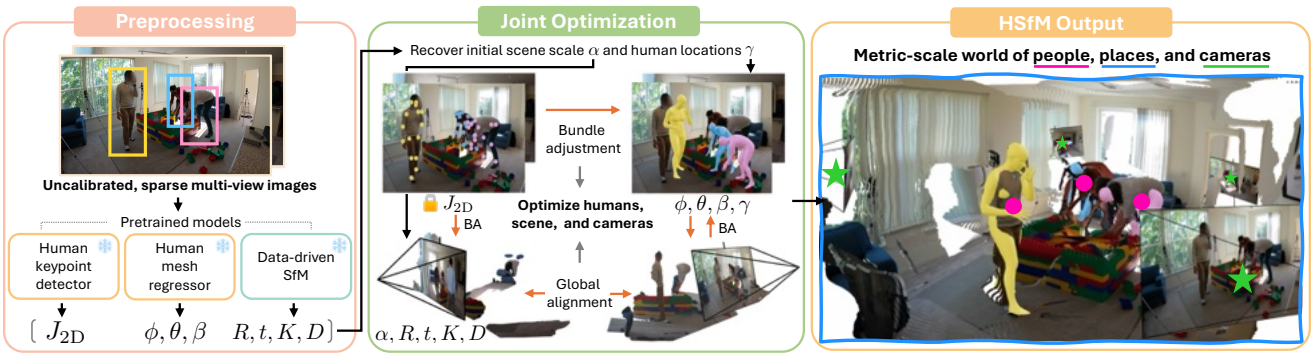


Figure 2. **Pipeline of Humans and Structure from Motion.** Our method processes synchronized images from an uncalibrated multi-view camera setup with known person correspondences across views. We utilize pretrained networks to estimate 2D human keypoints per image [59], 3D human mesh [16], scene point clouds in a *pointmap* representation, and camera intrinsic and extrinsic parameters [53]. We first initialize these estimates in a common world coordinate system by recovering the scene scale α and human locations (global translation in the world coordinate) γ, as described in Section 4.1. We then jointly optimize humans, the scene, and cameras using bundle adjustment based on 2D human keypoints, 3D human meshes, and a global alignment loss that merges per-view pointmaps into the same world space.

	Local Pose	Multi-Person	Stat. Scale	Camera	Places
HMR2 [16, 23]	✓	✗	✗	✗	✗
Multi-HMR [2]	✓	✓	✓	✗	✗
SLAHMR [61]	✓	✓	✓	✓	✗
UnCaliPose [55]	✓	✓	✓	✓	✗
DUSt3R [53]	✗	✗	✗	✓	✓
MASt3R [26]	✗	✗	✓	✓	✓
HSfM	✓	✓	✓	✓	✓

Table 1. **Comparison of methods across different features**. Previous works in human pose estimation like HMR2, SLAHMR, Multi-HMR, and UnCaliPose has made great progress in reconstructing body poses in single- and multi-person setups from images. Recent methods like DUSt3R and MASt3R are able to recover accurate camera poses and scene points clouds (places). This includes reconstructions with statistically correct scale (Stat. Scale) which can be obtained from human body models, *e.g.* in SLAHMR, or, as in MASt3R, from world knowledge. Our approach, HSfM, is the first to jointly reconstruct multiple people, scene, and cameras from sparse uncalibrated multi-view images.

ods [1, 9, 10], such as COLMAP [41, 42] employs keypoint detection, matching based on locally invariant descriptors [4, 30, 40], and incremental bundle adjustment to estimate camera poses and sparse 3D points. However, these traditional approaches are highly sensitive to noise at each stage of their sequential pipeline and require specific conditions for input, such as a large number of camera views with substantial overlapping image areas.

Learning-based SfM methods replace one or more components of the traditional SfM pipeline with data-driven approaches [3, 11, 28, 39, 62]. Recently, dense matching-based SfM [7, 21, 45, 50, 51, 70] has shifted from sparse keypoints to dense, data-driven approaches. DUSt3R [53] and more recent works [13, 26] exemplify this by predicting dense 3D pointmaps without requiring camera calibration. However, as these methods primarily focus on scene structure, they do not estimate human poses and face challenges in reconstructing pixels corresponding to people. In contrast, our approach robustly recovers both human poses and scene structure simultaneously.

SfM with Humans. Several recent works have leveraged humans in the scene as cues to overcome the limitations of traditional Structure from Motion (SfM) methods in challenging scenarios with minimal overlap or wide baselines. Ma et al. [32, 36] introduce the concept of Virtual Correspondences (VCs), which are pairs of pixels from different images whose camera rays intersect in 3D space, even if the points are not co-visible. Similarly, Xu et al. [58] addressed wide-baseline multi-camera calibration by employing 2D keypoint associations from people across different cameras, obtained by person re-identification methods. Xu and Kitani [56] extended this work by sequentially solving for person re-identification, camera pose estimation, and 3D human pose estimation using multi-view geometry and bundle adjustment in an optimization pipeline. While our approach aligns with this line of research, it differs by jointly optimizing people, places (scene pointmaps), and cameras.

Multi-view human reconstruction. In controlled environments with known camera parameters, multi-view reconstruction leverages geometric consistency for accurate 3D pose estimation, reducing single- and multi-person tasks to a triangulation problem [49] with a long history of research. Recent works explore setups with unknown camera poses by employing end-to-end learning methods that jointly estimate camera parameters and 3D poses [18, 63]. However, these methods are often limited to single-person scenarios [63] or do not incorporate scene context [18]. Existing multi-person methods focus on re-identification [6, 12, 22] or, for video, on re-identification and tracking [20].

In contrast to previous work, our approach does not require camera calibration. Instead, we leverage the human body structure and data-driven SfM methods to achieve accurate human pose and camera estimates.

3. Preliminaries and Notation

Setup. Our method takes as input an uncalibrated, sparse set of C images capturing people in a scene at a single moment in time. We denote each image as I^c, $c = \{1 \dots C\}$, corresponding to each camera, with resolution $H^c \times W^c$. We assume humans have been associated across views. Given this input, our method jointly reconstructs humans, scene, and cameras in a metric 3D world.

Human. For all of the following, we represent humans in the scene via a human body model, SMPL-X [35]. SMPL-X is a differentiable function that maps pose, $\theta \in \mathrm{SO}(3)^J$, and shape, $\beta \in \mathbb{R}^B$ to a triangulated mesh with J joints. This mesh can be placed in the world via two additional parameters, orientation, $\phi \in \mathrm{SO}(3)$, and translation, $\gamma \in \mathbb{R}^3$. We model multiple people, *i.e.* $h \in \{1 \dots H\}$ humans. In summary, a human, h, in the world is defined via

$$H^h = \{\phi^h, \theta^h, \beta^h, \gamma^h\}. \tag{1}$$

Cameras. To project 3D points onto an image, $I \in \mathbb{R}^{H \times W \times 3}$, we use a perspective camera model with intrinsics, $K \in \mathbb{R}^{3\times3}$ with focal lengths (f_x, f_y) and principal point $(W/2, H/2)$, and extrinsics with rotation, $R \in \mathrm{SO}(3)$, and translation, $t \in \mathbb{R}^3$. Existing methods produce camera estimates that are not necessarily scaled to real-world size. To address this, we introduce a scaling parameter, α, which adjusts the distance between cameras while preserving their relative directions. With these parameters a 3D point, x^{3D} can be projected to 2D via

$$x_{\text{2D}} = K(Rx_{\text{3D}} + \alpha t). \tag{2}$$

The pixel coordinates, $(u', v') = (\frac{u}{w}, \frac{v}{w})$ are obtained from $x_{\text{2D}} = [u, v, w]^\top$. K^c and R^c/t^c denote the in- and extrinsics of camera $c \in \{1, \dots C\}$.

Scene. We represent the scene via per-view pointmaps [53], $\mathcal{S} \in \mathbb{R}^{W \times H \times 3}$, a dense pixel-aligned 3D location for its corresponding image I in the world coordinate frame. S^c denotes the pointmap of an image c. A nice property of pointmap formulations is that we can express them through camera estimates and depth maps. For an image pixel (i, j) , its corresponding pointmap's world coordinate can be written as

$$\mathcal{S}_{i,j} = \alpha(R^\top[K^{-1}D_{i,j}[i, j, 1]^\top] - R^\top t). \tag{3}$$

This formulation unprojects a pixel (i, j) using its depth value $D_{i,j}$ and K, and maps it to the world coordinate system through $R^\top$, $-R^\top t$, *i.e.* the *camera-to-world* transformation defined by R and t, and scaling.

4. Humans and Structure from Motion

Our method takes as input an uncalibrated, sparse set of images capturing people in a scene at a single moment in time. Given this input, our goal is to jointly estimate each person's human parameters, the scene, and the camera parameters. Our key insight is that jointly reasoning about people, scene structure, and cameras improves all three aspects of reconstruction. To achieve this, we integrate global scene optimization from recent scene reconstruction methods with the traditional Structure-from-Motion (SfM) formulation. This integration leverages 2D human keypoints as reliable correspondences and 3D human meshes as robust 3D structures for bundle adjustment. Please also refer to Figure 2.

Our joint optimization approach has several advantages. By incorporating human mesh predictions, the method introduces metric scale into the reconstruction process, leveraging statistical information about human body dimensions. The cameras and 2D human keypoints enable precise positioning of individuals within the world coordinate system, allowing for the recovery of their heights and relative distances. Additionally, correspondences of 2D human keypoints enhance camera calibration, which in turn improves scene reconstruction. The scene structure further stabilizes the camera pose registration, creating a feedback loop that refines the overall system. The result is a globally consistent reconstruction of humans, scenes, and cameras, providing a comprehensive understanding of the environment. Since this is an under-determined problem, we take advantage of data-driven 3D human [16] and scene [53] reconstruction methods to provide initializations. Note that our method can easily integrate other mesh regressors tailored towards estimating non-standard body size such as BEV [46].

4.1. Initialization of World

The initial estimates of humans, scene structure, and cameras are derived from different networks and therefore exist in separate coordinate systems. Our objective is to align these components within a unified world coordinate system, a process we refer to as the initialization of the world. To achieve this, we estimate the metric scale that aligns the scene pointmaps and cameras with the humans.

One may simply start the optimization by setting the scale $\alpha = 1$. However, since SfM reconstructions are up-to-scale, the magnitude of t may vary significantly in every reconstruction. If the SfM scene is too small relative to people, cameras may be in-front of the humans, *e.g.* when the scene is placed inside the human mesh, leading to degenerate solutions. Setting alpha to arbitrary big value can prevent this, but makes the problem prone to local minima.

We propose an analytical method to approximate a consistent reconstruction using data-driven outputs of 2D/3D human keypoints and camera parameters. Initially, we roughly position individuals within the scene by estimat-

ing γ, based on the predicted 2D/3D human keypoints and the focal length from $\tilde{K}$, following a similar approach to Ye et al. [61]. Next, we calculate the initial scale α of the data-driven SfM-predicted world by aligning the data-driven with human-centric camera positions.

Specifically, we first obtain each camera's rotation, $\hat{R}^c$, using the estimated 3D human body orientation, $\tilde{\phi}$. This leverages the fact that the human's orientation should remain consistent in the world coordinate frame across different views. Assuming a reference camera, c_1, and an anchor person, $\tilde{\phi}^h$, we recover each camera's rotation by solving

$$(\hat{R}^c)^\top = (R^{c_1})^\top \tilde{\phi}^{hc_1} (\tilde{\phi}^{hc})^\top. \tag{4}$$

We pick the anchor person h based on the best view coverage w.r.t. to the 2D joint confidence scores.

We estimate the camera translation by first estimating the location of the person γ in the world using the data-driven focal length prediction from $\tilde{K}$ and the size of the predicted human following the similar triangle ratio using the average 2D and 3D bone lengths as done in Ye *et al*. [61]. Because the human position in the world coordinate frame should remain consistent when viewed from any camera, we recover camera position T^c in the world coordinate frame via

$$\hat{T}^c = \tilde{\gamma}^{c_1} - (\hat{R}^c)^\top \tilde{\gamma}^c. \tag{5}$$

Finally, given the camera positions $\hat{T}$ derived from the humans and the data-driven camera position predictions $\tilde{T}$, we solve a least-squares problem to compute $\hat{\alpha}$, which aligns the scene pointmap $\mathcal{S}$ and the camera translation t with the metric-scale world defined by the humans $\tilde{H}$. The scaling factor $\hat{\alpha}$ provides a reasonable approximation of the metric scale. For the initial estimates of $\tilde{H}$, we rely on an estimate from a reference camera in practice. Please see the project page supplementary video for an animated illustration.

4.2. Reconstructing People, Places, and Cameras

After initializing the world, we jointly optimize the humans, depth maps, and cameras using a global scene optimization loss and bundle adjustment guided by 2D human keypoint predictions. The objective function is defined as follows:

$$\min_{\{\alpha,\gamma,\beta,\phi,\theta,R,t,K,D\}} L_{\text{Humans}} + \lambda L_{\text{Places}}. \tag{6}$$

To gradually guide the optimization towards the global minimum, we first optimize $\{\alpha, \gamma, \beta\}$ with $\lambda = 0$. Then, we set λ and optimize $\{\gamma, \beta, \phi, \theta, R, t, K, D\}$. The result of this optimization is a metric-scale world with consistent humans, scene, and cameras.

Bundle adjustment based on human keypoints: We define the bundle adjustment objective as follows:

$$L_{\text{Humans}} = \frac{1}{HC} \sum_{c=1}^{C} \sum_{h=1}^{H} L_J^{ch} + \frac{1}{H} \sum_{h=1}^{H} L_\beta^h. \tag{7}$$

The term L_J denotes the re-projection error between the 3D joints of the current estimate with the estimated camera and detected 2D keypoints by ViTPose [59]:

$$L_J^{ch} = \frac{1}{b_{\text{2D}}^{ch}} ||c_{\text{2D}}^{ch} (J_{\text{2D}}^{ch} - K^c (R^c J_{\text{3D}}^h + \alpha t^c))||_2 \tag{8}$$

The keypoint loss for each person, h, and camera, c, is normalized by the bounding box height b_{2D}^{ch} of the 2D human detection and detected keypoints weighted by their estimated confidence scores c_{2D}^{ch}. We further regularize human body shape, $\tilde{\beta}$, to stay close to the average shape of the human body model:

$$L_\beta^h = ||\beta^h||_2. \tag{9}$$

The output of this optimization is people in plausible size and locations in the world coordinate frame. Please refer to the supplementary material for implementation details.

Global scene optimization: We adapt the global alignment loss from DUSt3R [53], which originally optimizes camera and world pointmaps. Intuitively, this loss takes pairs of cameras, say A and B (c_i and c_j), with the predicted content of c_i in B's view. The alignment loss transforms this "A's-content-in-B's-view" to the world coordinate frame where it's compared against the optimized global scene pointmap, which is A's content in the world.

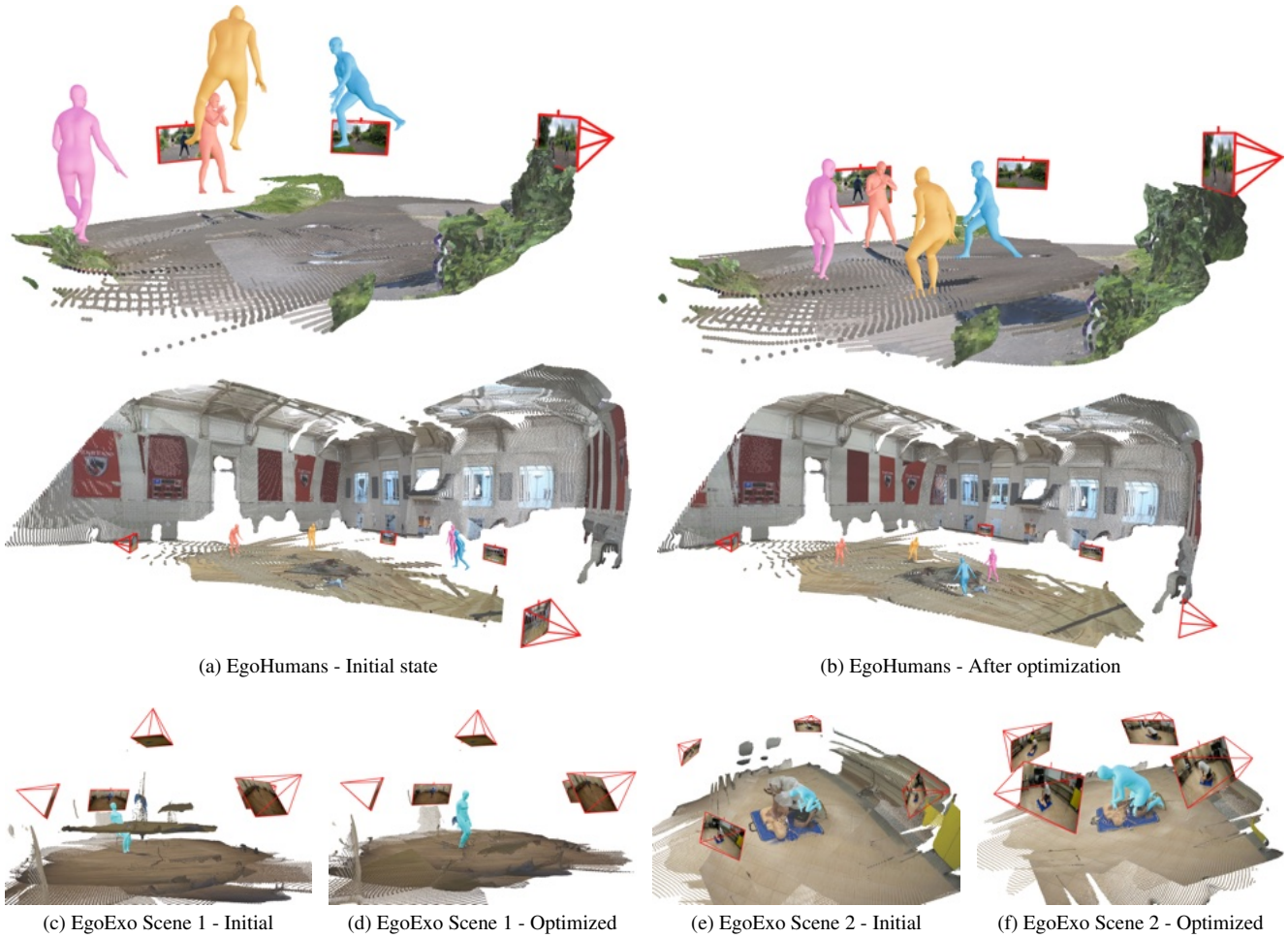
More formally, the global alignment loss aligns per-view pointmaps into a joint world space, *i.e.* $\{\mathcal{S}^c \in \mathbf{R}^{H \times W \times 3}\}$ for $c = 1 \dots C$. To achieve this, the alignment loss takes cross-view pointmaps X^{c_i,c_j} for two cameras $(c_i, c_j) \in \mathcal{E}$, with $i, j = 1 \dots C$ and $\mathcal{E}$ being the set of all pairs of cameras with $i \neq j$. The notation, X^{c_i,c_j}, describes cross-view pointmaps, meaning that the pointmap X^{c_i} of camera c_i is expressed in the coordinate frame of camera c_j. Pairwise transformation matrices $P^{c_i,c_j \to w} \in \mathbb{R}^{3\times4}$, *i.e.* the transformation matrix for camera pair c_i, c_j that brings c_j's content to the world coordinate frame. The global alignment loss term is defined as

$$L_{\text{Places}} = \sum_{(c_i,c_j)\in\mathcal{E}} \sum_{i=1}^{HW} Q_i^{c_i,c_j} ||\mathcal{S}_i^{c_i} - \sigma^{c_i,c_j \to w} P^{c_i,c_j \to w} X_i^{c_i,c_j}||, \tag{10}$$

where $Q_i^{c_i,c_j}$ are predicted per-pixel confidence maps. $\sigma^{c_i,c_j \to w}$ is a scaling factor associated to the pair (c_i, c_j). Note that different from DUSt3R [53], we don't regularize it to avoid a trivial optimum, since the scale is constrained by humans. We omit σ and P in Eq. 6 for clarity, but they are still optimized together following DUSt3R.

5. Experiments

We evaluate HSfM's effectiveness in terms of human pose estimation within the world coordinate system and camera



(a) EgoHumans - Initial state

(b) EgoHumans - After optimization

(c) EgoExo Scene 1 - Initial

(d) EgoExo Scene 1 - Optimized

(e) EgoExo Scene 2 - Initial

(f) EgoExo Scene 2 - Optimized

Figure 3. **Qualitative results from HSfM.** We show our optimized result on sequences from EgoHumans (top) and EgoExo4D (bottom). Note how in the Initial state (left) people are floating in the air (a), how the scene and human scale is not aligned (e), and how noisy the scene appears (c). Our method resolves these problems by grounding people in the scene (b), recovering plausible metric scale (f), and better camera estimates (d). We achieve this without scene contact constraints, which often require assumptions about the environment—such as flat terrain—or about motion, such as the assumption that humans are always in contact with the ground (i.e., no jumping). For more qualitative results, including a demo on images taken in the wild with a minimal capturing setup, please see our supplementary material.

accuracy, and show qualitative results of our joint optimization on humans, scene pointmaps, and cameras.

Evaluation Datasets: We evaluate on EgoHumans [24] and EgoExo4D [17]. EgoHumans is a multi-view, multi-human benchmark for human pose estimation, featuring videos of 2-4 people in real-world activities. EgoExo4D is a large-scale dataset of people performing tasks like dancing, playing music, or bike repair; see Sup. Mat. for details.

Evaluation Metrics: We report metrics for humans and cameras. For people, we use the Mean Per-Joint Position Error (MPJPE). We report **W-MPJPE**, the metric measured in the world coordinate system, and **PA-MPJPE**, its Procrustes-Aligned version, measuring the local pose accuracy. We introduce Group-Aligned MPJPE, (**GA-MPJPE**), which evaluates the relative distance after Sim(3) alignment between people. All metrics are reported in meters.

For cameras, we report average camera translation error **TE**, *i.e.* the mean euclidean distance in meters between predicted and ground truth camera translations after SE(3) alignment. TE evaluates accuracy in the metric prediction. We also report the Sim(3) aligned version, **s-TE**. **AE** measures the camera Angle Error, *i.e.* the mean Euclidean distance between predicted and ground truth camera rotation. **RRA** [52] evaluates the Relative Rotation Accuracy by comparing the relative rotation between two predicted cameras with the corresponding ground truth. **CCA** [27] assesses the Camera Center Accuracy by directly comparing the predicted and ground truth camera poses. While Lin et al. [27] reported CCA only after optimal Sim(3) alignment, we provide results for two variants: the default

	Method	Human Metrics						Camera Metrics					
		W-MPJPE↓	GA-MPJPE↓	PA-MPJPE↓	TE↓	s-TE↓	AE↓	RRA@10↑	RRA@15↑	CCA@10↑	CCA@15↑	s-CCA@10↑	s-CCA@15↑
EgoHumans	UnCaliPose* [55]	3.51	0.67	0.13	2.63	2.63	60.90	0.28	0.39	-	-	0.33	0.44
	DUSt3R [53]	-	-	-	-	1.15	11.00	0.61	0.74	-	-	0.49	0.74
	MASt3R [26]	-	-	-	4.97	0.92	10.42	0.61	0.74	0.06	0.07	0.65	0.86
	HSfM (init.)	4.28	0.51	0.06	2.37	1.15	11.00	0.52	0.79	0.26	0.38	0.49	0.74
	HSfM (Ours)	**1.04**	**0.21**	**0.05**	**2.09**	**0.75**	**9.35**	**0.72**	**0.89**	**0.32**	**0.46**	**0.75**	**0.91**
EgoExo4D	UnCaliPose* [55]	3.59	-	1.19	2.21	0.98	63.98	0.20	0.31	-	-	0.26	0.37
	DUSt3R [53]	-	-	-	-	**0.34**	10.06	**0.81**	0.88	-	-	0.64	**0.84**
	MASt3R [26]	-	-	-	1.03	0.36	**9.11**	**0.81**	**0.90**	**0.09**	**0.17**	**0.70**	0.81
	HSfM (init.)	5.80	-	0.08	1.27	**0.34**	10.06	**0.81**	0.88	0.05	0.10	0.64	**0.84**
	HSfM (Ours)	**0.50**	-	**0.07**	**1.01**	**0.34**	10.39	0.80	0.89	0.05	0.14	**0.70**	**0.84**

Table 2. **Evaluation on EgoHumans and EgoExo4D.** HSfM outperforms existing human and scene reconstruction methods, delivering metric-scale reconstructions for humans, the scene, and cameras within the same world coordinate system. Our approach shows major improvement over the initial estimates HSfM (init.), obtained from DUSt3R [53] and HMR2 [16] (Section 4.1), particularly when multiple humans are present in the scene, as seen in EgoHumans, compared to a single individual in EgoExo4D. Additionally, HSfM surpasses MASt3R in metric-scale camera metrics, demonstrating the benefit of human size for recovering scene scale. All baselines and our method use four cameras on EgoHumans and up to six cameras on EgoExo4D. Human and camera translation metrics are reported in meters.

Method	Human Metrics						Camera Metrics					
	W-MPJPE↓	GA-MPJPE↓	PA-MPJPE↓	TE↓	s-TE↓	AE↓	RRA@10↑	RRA@15↑	CCA@10↑	CCA@15↑	s-CCA@10↑	s-CCA@15↑
M0: HSfM (init.)	4.28	0.51	0.06	2.37	1.15	11.00	0.52	0.79	0.26	0.38	0.49	0.74
M1: S & C w/o L_{Humans}	3.94	0.57	0.10	2.13	1.1	10.93	0.52	0.79	0.27	0.40	0.48	0.77
M2: w/o L_{Places}	1.29	0.24	0.05	2.82	0.87	13.02	0.50	0.73	0.16	0.24	0.72	0.88
M3: HSfM (Ours)	**1.04**	**0.21**	**0.05**	**2.09**	**0.75**	**9.35**	**0.72**	**0.89**	**0.32**	**0.46**	**0.75**	**0.91**
HSfM (init.)	3.87	0.70	**0.06**	1.88	1.15	11.34	0.47	0.77	0.23	0.37	0.38	0.69
1 Human	1.69	0.58	**0.06**	1.91	1.21	10.31	0.52	0.82	0.27	0.44	0.44	0.68
2 Humans	1.66	0.49	**0.06**	1.84	1.10	9.41	0.62	0.87	0.33	0.45	0.53	0.75
3 Humans	1.41	**0.38**	**0.06**	1.69	0.93	8.21	**0.74**	**0.92**	0.32	0.46	0.58	0.78
4 Humans	**1.28**	0.39	**0.06**	**1.52**	**0.77**	**8.11**	**0.74**	0.90	**0.34**	**0.53**	**0.73**	**0.90**

Table 3. **Ablation study**. We demonstrate the advantages of our joint optimization by removing each terms on the EgoHumans dataset. The results indicate that joint optimization is crucial for achieving a coherent reconstruction of cameras and humans. We also perform an ablation study to investigate how the number of humans included affects both camera and human pose estimation in the world coordinates by varying the number of people to include in the optimization using a subset of EgoHumans containing scenes with four people. The results reveal that the effect scales with the number of humans, highlighting the importance of leveraging multiple individuals.

metric computed after SE(3) alignment, and **s-CCA**, computed after Sim(3) alignment. RRA, CCA, and s-CCA are reported for a threshold τ following the previous literature [13, 27, 52, 53]. Note that W-MPJPE, TE, and CCA evaluate absolute Euclidean error, while GA-MPJPE, PA-MPJPE, s-TE, and s-CCA evaluate errors up to scale. For further details, please refer to the supplementary material.

5.1. Results

We compare human and camera estimation baselines on EgoHumans and EgoExo4D (Tab. 2), including UnCaliPose [55], DUSt3R [53], and MASt3R [26]. UnCaliPose jointly reconstructs humans and cameras using SfM but does not reconstruct the scene and relies on ground truth bone lengths during testing. For a fair comparison, we assume known re-identification across views, use ViTPose for UnCaliPose and HSfM, and apply DUSt3R's focal length to UnCaliPose. Since no existing approach jointly estimates people, places, and cameras at metric scale from sparse multi-view images, we also report HSfM (init.), the state after our initialization in Sec. 4.1.

On EgoHumans, HSfM (init.) outperforms UnCaliPose in scale-normalized human metrics, reducing GA-MPJPE by approximately 24% and PA-MPJPE by over 50%. It nearly doubles relative rotation accuracy compared to UnCaliPose with RRA@10 and RRA@15 improvements of 86% and 100%, respectively. These gains show the strength of leveraging HMR2 and DUSt3R for initialization.

Our optimization further improves results: W-MPJPE drops from 4.28m in HSfM (init.) to 1.04m in HSfM, demonstrating the effectiveness of our approach in resolving scale ambiguity and accurately positioning humans within the world. The camera metrics also improve substantially, surpassing DUSt3R's initial outputs and outperforming MASt3R, reducing TE from 4.97m to 2.09m and achieving about seven times better CCA@15. These results

highlight the advantages of incorporating humans into the reconstruction process to achieve a consistent metric-scale world, consistent with the findings of Zhao et al. [68].

Similar trends observed in EgoExo4D further validate the effectiveness of our method. HSfM achieves substantially better human metrics compared to UnCaliPose, reducing W-MPJPE from 3.59m to 0.50m and PA-MPJPE from 0.13m to 0.07m. For metric-scale camera metrics, HSfM outperforms HSfM (init.), improving CCA@15 by 33%. The improvements in scale-invariant camera metrics are smaller than on EgoHumans, likely due to heavy indoor occlusions affecting 2D keypoint predictions and the use of only one person in EgoExo4D. Similarly, MASt3R estimates slightly more accurate camera centers (CCA@10/15) on EgoExo4D while HSfM is on-par and slightly better on the scale-invariant version (s-CCA@10/15). This is likely due to a single person being less effective for estimating scene scale; an effect we ablate in Tab. 3 on EgoHumans where we also observe a single person to be less efficient for estimating scene scale compared to more people. In contrast, EgoHumans benefits from multiple individuals, providing more 2D keypoint correspondences and strengthening optimization. Our ablation study confirms this trend.

Ablations: We validate the importance of jointly optimizing humans, scenes, and cameras in Table 3. The first variant, M1, detaches the gradients from the human loss to the scene and camera parameters while still optimizing all parameters. Essentially, the cameras and scene do not adjust to minimize the human losses. This leads to minor W-MPJPE improvement (4.28m to 3.94m) and slightly higher GA-MPJPE (0.51m to 0.57m). Camera metrics nearly stagnate (RRA@15, CCA@15) since human losses do not influence the optimization of camera and scene. The second variant, M2, optimizes cameras and humans solely based on the human loss, excluding the scene loss. Interestingly, this significantly improves W-MPJPE (4.28m to 1.29m) and GA-MPJPE (0.57m to 0.24m), indicating accurate recovery of human world locations and relative distances. However, the camera metrics degrade considerably: CCA@15 by 36.8%, and RRA@15 by 7.6%. Without scene losses, the structure fails to anchor the cameras and overfits to human keypoints. This behavior is similar to the limitations observed in UnCaliPose, which relies solely on human keypoints for SfM. In contrast, our full method (M3) achieves the best metrics, reducing W-MPJPE and GA-MPJPE to 1.04m and 0.21m, respectively, and increasing RRA@15 by 22% and CCA@15 by 21%. This highlights the importance of jointly optimizing humans, scenes, and cameras to achieve a coherent and accurate metric-scale reconstruction.

We conduct an ablation study to investigate the impact of the number of humans on camera estimation and its subsequent effect on human metrics in the world coordinate system. As shown in the table, adding more humans consistently improves camera pose estimation, particularly by reducing camera translation error. This improvement directly contributes to achieving the lowest W-MPJPE, accurately reconstructing human locations in the world.

The results highlight that increasing the number of humans, *i.e.* introducing more correspondences for Structure-from-Motion (SfM), strengthens the bundle adjustment process. This validates our strategy of integrating global scene optimization techniques from recent scene reconstruction methods with the traditional SfM formulation. Our approach effectively leverages 2D human keypoints as reliable correspondences and 3D human meshes as robust structures to enhance bundle adjustment and produce coherent, metric-scale reconstructions.

Please refer to the supplementary material for an ablation study on camera/scene scale initialization and the impact of the number of cameras.

Qualitative Results: See Figs. 1 and 3 for qualitative results on EgoHumans and EgoExo4D and our Sup. Mat. for an *in-the-wild* demo with images we captured using a minimal setup of just two cell phones. Figure 3 shows intermediate optimization steps, where in the beginning, people are floating around in mid air. After joint optimization, their feet are consistent with the environment, without any explicit contact constraints. The structure also improves as pointmaps are more coherent around the ground. Please note that the human point cloud is not used as prior for human reconstruction, *i.e.* noisy point clouds or gaps do not directly affect the human reconstruction accuracy. Despite clear improvements in human localization, we sometimes observe scenes with slightly uneven ground planes indicating that scene reconstruction remains a challenging problem. Nonetheless, our approach improves camera metrics and scene reconstruction qualitatively (see Fig. S.3).

6. Conclusion

In this work, we propose Humans and Structure from Motion, HSfM, which optimizes humans, cameras, and scenes in a joint framework. We build on the success of data-driven learning in two parallel domains – 2D/3D human reconstruction [16, 59] and scene reconstruction [53]. However, neither of these approaches is able to reconstruct humans, scenes, and cameras coherently. Our experiments verify the synergy between these three elements — integrating human reconstruction into the classic SfM task not only properly places people in the world, but also significantly improves camera pose accuracy. Despite promising results, as optimization-based framework our approach can be sensitive to hyper-parameters. In future work, it would be interesting to explore this synergy in a feed-forward framework with integrated re-ID or leverage recent work like MONSt3R [66] to extend our insights to videos.

Acknowledgements

This project is supported in part by DARPA No. HR001123C0021, IARPA DOI/IBC No. 140D0423C0035, NSF:CNS-2235013, ONR MURI N00014-21-1-2801, Bakar Fellows, and Bair Sponsors. The views and conclusions contained herein are those of the authors and do not represent the official policies or endorsements of these institutions. We also thank Chung Min Kim for her critical reviews on this paper and Junyi Zhang for his valuable insights on the method.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *ACM Communications*, 2011. 3

[2] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. *arXiv preprint arXiv:2402.14654*, 2024. 3, 7

[3] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.net: Keypoint detection by handcrafted and learned cnn filters. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, 2006. 3

[5] Paul Beardsley, Phil Torr, and Andrew Zisserman. 3d model acquisition from extended image sequences. In *European Conference on Computer Vision (ECCV)*, 1996. 2

[6] Chao Chen, Georgios Pavlakos, Shihao Zou, Tony Tung, and Katerina Fragkiadaki. Multi-person 3d pose estimation in crowded scenes based on multi-view geometry. *arXiv preprint arXiv:2007.10986*, 2020. 3

[7] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[8] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[9] David Crandall, Andrew Owens, Noah Snavely, and Daniel Huttenlocher. Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2013. 3

[10] Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu. Hsfm: Hybrid structure-from-motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Workshops*, 2018. 3

[12] Junting Dong, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7792–7801, 2019. 2, 3

[13] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: A fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv*, 2024. 3, 7

[14] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 7

[15] Andrew W Fitzgibbon and Andrew Zisserman. Automatic camera recovery for closed or open image sequences. In *European Conference on Computer Vision (ECCV)*, 1998. 2

[16] 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, 7, 8, 1

[17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 6, 1

[18] Peter Hardy and Hansung Kim. Unsupervised multi-person 3d human pose estimation from 2d poses alone. *arXiv preprint arXiv:2309.14865*, 2023. 3

[19] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. 2004. 2

[20] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 13264–13275, 2022. 2, 3

[21] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[22] Amir Kadkhodamohammadi and Nicolas Padoy. A generalizable approach for multi-view 3d human pose regression. *arXiv preprint arXiv:1804.10462*, 2018. 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)*, pages 7122–7131, 2018. 2, 3, 7

[24] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark. *arXiv preprint arXiv:2305.16487*, 2023. 2, 6, 1

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 7

[26] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 7

[27] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 6, 7, 4

[28] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[29] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M Rehg, and Siyu Tang. 4d human body capture from egocentric video via 3d scene grounding. In *3DV*, 2021. 7

[30] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 2004. 3

[31] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 1

[32] Wei-Chiu Ma, Alexander J. Yang, Shenlong Wang, Raquel Urtasun, and Antonio Torralba. Virtual correspondence: Humans as a cue for extreme-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[33] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael J. Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3D social interaction from images. 2024. 7

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[35] 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)*, pages 10975–10985, 2019. 4, 7

[36] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1485–1495, 2022. 3

[37] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[38] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[39] Jérôme Revaud, César Roberto de Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[40] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. Orb: An efficient alternative to sift or surf. In *International Conference on Computer Vision (ICCV)*, 2011. 3

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

[42] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 3

[43] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 7

[44] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 7

[45] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[46] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13243–13252, 2022. 4, 7

[47] 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/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 7

[48] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9:137–154, 1992. 2

[49] Bill Triggs, Paul F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice*, pages 298–372. Springer, Berlin, Heidelberg, 2000. 3

[50] Prune Truong, Martin Danelljan, and Radu Timofte. Glunet: Global-local universal network for dense flow and correspondences. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[51] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[52] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *International Conference on Computer Vision (ICCV)*, pages 9773–9783, 2023. 6, 7, 4

[53] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 2, 3, 4, 5, 7, 8, 1

[54] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. *arXiv preprint arXiv:2403.17346*, 2024. 7

[55] Yan Xu and Kris Kitani. Multi-view multi-person 3d pose estimation with uncalibrated camera networks. In *British Machine Vision Conference (BMVC)*, 2022. 2, 3, 7

[56] Yan Xu and Kris Kitani. Multi-view multi-person 3d pose estimation with uncalibrated camera networks. In *British Machine Vision Conference (BMVC)*, 2022. 2, 3

[57] 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. 1

[58] Yan Xu, Yu-Jhe Li, Xinshuo Weng, and Kris Kitani. Wide-baseline multi-camera calibration using person re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[59] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 2, 3, 5, 8, 1

[60] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 2

[61] 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)*, 2023. 3, 5, 1, 2, 7

[62] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[63] Tao Yu, Zerong Zheng, Kaiwen Guo, and Yebin Liu. Multi-view human body reconstruction from uncalibrated cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[64] Yuming Yuan, Chun-Hao P. Huang, Donglai Xiang, Yufeng Zhou, Mengcheng Xu, Jingwei Huang, Chenxi Jiang, Tzu-Mao Xu, Deva Ramanan, and Michael J. Black. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[65] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022. 7

[66] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *ICLR*, 2025. 8

[67] 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)*, pages 180–200, 2022. 7

[68] Yizhou Zhao, Hengwei Bian, Kaihua Chen, Pengliang Ji, Liao Qu, Shao-yu Lin, Weichen Yu, Haoran Li, Hao Chen, Jun Shen, et al. Metric from human: Zero-shot monocular metric depth estimation via test-time adaptation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 8

[69] Yizhou Zhao, Tuanfeng Y. Wang, Bhiksha Raj, Min Xu, Jimei Yang, and Chun-Hao P. Huang. Synergistic global-space camera and human reconstruction from videos. 2024. 1, 7

[70] Shengjie Zhu and Xiaoming Liu. Pmatch: Paired masked image modeling for dense geometric matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[71] Yuliang Zou, Jimei Yang, Duygu Ceylan, Jianming Zhang, Federico Perazzi, and Jia-Bin Huang. Reducing footskate in human motion reconstruction with ground contact constraints. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 2