SLOPER4D: A Scene-Aware Dataset for Global 4D Human Pose Estimation in Urban Environments

Yudi Dai[1,2] Yitai Lin[1,2] Xiping Lin[1,2] Chenglu Wen[1,2*] Lan Xu[3]
Hongwei Yi[4] Siqi Shen[1,2] Yuexin Ma[3] Cheng Wang[1,2]

[1] Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University
[2] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University
[3]ShanghaiTech University, China [4]Max Planck Institute for Intelligent Systems, Germany

Figure 1. Using the head-mounted LiDAR and camera to scan the IMUs wearer, we construct SLOPER4D, a large scene-aware dataset for global 4D human pose estimation in urban environments.

Abstract

We present SLOPER4D, a novel scene-aware dataset collected in large urban environments to facilitate the research of global human pose estimation (GHPE) with human-scene interaction in the wild. Employing a head-mounted device integrated with a LiDAR and camera, we record 12 human subjects' activities over 10 diverse urban scenes from an egocentric view. Frame-wise annotations for 2D key points, 3D pose parameters, and global translations are provided, together with reconstructed scene point clouds. To obtain accurate 3D ground truth in such large dynamic scenes, we propose a joint optimization method to fit local SMPL meshes to the scene and fine-tune the camera calibration during dynamic motions frame by frame, resulting in plausible and scene-natural 3D human poses. Eventually, SLOPER4D consists of 15 sequences of human motions, each of which has a trajectory length of more than 200 meters (up to 1,300 meters) and covers an area of more than 200 m^2 (up to 30,000 m^2), including more than 100k LiDAR frames, 300k video frames, and 500k IMU-based motion frames. With SLOPER4D, we provide a detailed and thorough analysis of two critical tasks, including camera-based 3D HPE and LiDAR-based 3D HPE in urban environments, and benchmark a new task, GHPE. The in-depth analysis demonstrates SLOPER4D poses significant challenges to existing methods and produces great research opportunities. The dataset and code are released at http://www.lidarhumanmotion.net/sloper4d/.

1. Introduction

Urban-level human motion capture is attracting more and more attention, which targets acquiring consecutive fine-grained human pose representations, such as 3D skeletons

*Corresponding author.

and parametric mesh models, with accurate global locations in the physical world. It is essential for human action recognition, social-behavioral analysis, and scene perception and further benefits many downstream applications, including Augmented/Virtual Reality, simulation, autonomous driving, smart city, sociology, etc. However, capturing extra large-scale dynamic scenes and annotating detailed 3D representations for humans with diverse poses is not trivial.

Over the past decades, a large number of datasets and benchmarks have been proposed and have greatly promoted the research in 3D human pose estimation (HPE). They can be divided into two main categories according to the capture environment. The first class usually leverages marker-based systems [16, 33, 45], cameras [14, 59, 60], or RGB-D sensors [13, 64] to capture human local poses in constrained environments. However, the optical system is sensitive to light and lacks depth information, making it unstable in outdoor scenes and difficult to provide global translations, and the RGB-D sensor has limited range and could not work outdoors. The second class [39, 49] attempts to take advantage of body-mounted IMUs to capture occlusion-free 3D poses in free environments. However, IMUs suffer from severe drift for long-term capturing, resulting in misalignments with the human body. Then, some methods exploit additional sensors, such as RGB camera [17], RGB-D camera [46, 57, 67], or LiDAR [27] to alleviate the problem and make obvious improvement. However, they all focus on HPE without considering the scene constraints, which are limited in reconstructing human-scene integrated digital urban and human-scene natural interactions.

To capture human pose and related static scenes simultaneously, some studies use wearable IMUs and body-mounted camera [12] or LiDAR [5] to register the human in large real scenarios and they are promising for capturing human-involved real-world scenes. However, human pose and scene are decoupled in these works due to the ego view, where auxiliary visual sensors are used for collecting the scene data while IMUs are utilized for obtaining the 3D pose. Different from them, we propose a novel setting for human-scene capture with wearable IMUs and global-view LiDAR and camera, which can provide multi-modal data for more accurate 3D HPE.

In this paper, we propose a huge scene-aware dataset for sequential human pose estimation in urban environments, named SLOPER4D. To our knowledge, it is the first urban-level 3D HPE dataset with multi-modal capture data, including calibrated and synchronized IMU measurements, LiDAR point clouds, and images for each subject. Moreover, the dataset provides rich annotations, including 3D poses, SMPL [32] models and locations in the world coordinate system, 2D poses and bounding boxes in the image coordinate system, and reconstructed 3D scene mesh. In particular, we propose a joint optimization method for obtaining accurate and natural human motion representations by utilizing multi-sensor complementation and scene constraints, which also benefit global localization and camera calibration in the dynamic acquisition process. Furthermore, SLOPER4D consists of over 15 sequences in 10 scenes, including library, commercial street, coastal runway, football field, landscape garden, etc., with up to 30k m^2 area size and $200 \sim 1,000m$ trajectory length for each sequence. By providing multi-modal capture data and diverse human-scene-related annotations, SLOPER4D opens a new door to benchmark urban-level HPE.

We conduct extensive experiments to show the superiority of our joint optimization approach for acquiring high-quality 3D pose annotations. Additionally, based on our proposed new dataset, we benchmark two critical tasks: camera-based 3D HPE and LiDAR-based 3D HPE, as well as provide benchmarks for GHPE.

Our contributions are summarized as follows:

- We propose the first large-scale urban-level human pose dataset with multi-modal capture data and rich human-scene annotations.
- We propose an effective joint optimization method for acquiring accurate human motions in both local and global by integrating LiDAR SLAM results, IMU poses, and scene constraints.
- We benchmark two HPE tasks as well as a GHPE task on SLOPER4D, demonstrating its potential of promoting urban-level 3D HPE research.

2. Related Work

2.1. 3D Human Motion Datasets

Many datasets have been proposed with different sensors and setups to facilitate the research on 3D human pose estimation. The H3.6M [16] is a large-size dataset providing synchronized video with optical-based MoCap in studio environments. To perform markerless capture in different indoor scenes, PROX [13] uses an RGB-D sensor to scan a single person. EgoBody [64] uses multiple RGB-D sensors to pre-scan the room and scan the interacting persons. LiDARHuman26M [27] can capture long-range human motions with static LiDAR and IMUs. However, they are limited to static environments, human activities, and interactions. 3DPW [49] is the first dataset providing 3D annotations in the wild which uses a single hand-held RGB camera to optimize human pose from IMUs for a certain period of frames. It doesn't provide accurate global translation and 3D scenes. HPS [12] reconstructs the human body pose using IMUs and self-localizes it with a head-mounted camera in large 3D scenes, but it heavily relies on the pre-built map. HSC4D [5] removes the reliance on the pre-built map and achieves global human motion capture in large scenes.

Dataset	In the wild	Global	3D Scene	Point cloud	Video	IMU	# Scene	# Area size (m^2)	# Subject	# Frame
H3.6M [16]	✗	✓	✗	✗	✓	✗	-	12	11	3.6M
3DPW [49]	✓	✗	✗	✗	✓	✓	-	< 300	7	51k
PROX [13]	✗	✓	✓	✗	✓	✗	12	< 30	20	20k
HPS [12]	✓	✓	✓	✗	✓*	✓	8	$< 1{,}000$	7	7k
HSC4D [5]	✓	✓	✓	✓*	✗	✓	5	$< 5{,}000$	2	10k
LH26M [27]	✗	✗	✗	✓	✓	✓	-	< 200	13	184k
EgoBody [64]	✗	✓	✓	✗	✓	✗	15	< 50	20	153k
SLOPER4D	✓	✓	✓	✓	✓	✓	10	$< 30{,}000$	12	100k

Table 1. **Comparisons with existing datasets.** "Global" denotes to human poses with **global translation**. The "area size" is estimated with the published data. The * indicates the data modality is only used for human self-localization rather than for human-related data.

However, the camera in HPS and the LiDAR in HSC4D are only used to perceive the environment rather than capture human data. With the scene-aware dataset we proposed for global human pose estimation, we can benchmark the 3D HPE in the wild with the LiDAR or camera modalities.

2.2. Human Localization and Scene Mapping

Human self-localization aims at estimating the 6-DoF of the human subject in global coordinates. The image-based methods [21, 37, 52] regress locations directly from a single image with a pre-built map. The scene-specific property makes them hard to generalize to unseen scenes. LiDAR is widely used in Simultaneous Localization and Mapping (SLAM) [4, 28, 41, 62] due to its robustness and low drift. To address the drift problem and improve robustness in dynamic motions, RGB cameras [40, 44, 63], IMU [10, 36, 42], or both [6, 43, 68], have been integrated with the mapping task. Most attention has been paid to autonomous driving [9] [23] or robotics from the third-person view and they usually do not focus on humans. To achieve self-localization, LiDAR is designed as backpacked [20, 31, 54] and handheld [2]. To efficiently capture human motions and reconstruct urban scenes, we utilize LiDAR with a built-in IMU (different from the IMUs for motion capture) and propose a pipeline for constructing multi-modal data. This approach provides accurate information on human motions at both local and global levels, as well as enables mapping in large outdoor environments.

2.3. Global 3D Human Pose Estimation

Most studies recover human meshes in camera coordinate [26, 66] or root-relative poses [18, 24, 25]. Recovering global human motions in unconstrained scenes is a challenging topic in computer vision and has gained more and more research interest in recent years. IMU sensors are widely used in commercial [38, 39] and research activities [15, 48, 51], and are attached to body limbs to capture human motions in studio-environments. But it suffers severe drift in the wild. Some methods rely on additional RGB [8, 34, 49, 50] or pre-scan maps [12], or LiDAR [5] to complement the IMUs in large-scale scenes. Based on human-scene interaction, some work proposed scene-aware solutions using static cameras [14, 60] to obtain accurate and scene-natural human motions. 4DCapture [30] uses a dynamic head-mounted camera to self-localize and reconstruct the scene with the Struct From Motion method. However, it often fails when the illumination changes in the wild. MOVER [60] uses a single camera to optimize the 3D objects in a static scene, resulting in better 3D scene reconstruction and human motions. GLAMR [61] uses global trajectory predictions to constrain both human motions and dynamic camera poses, achieving state-of-the-art results on in-the-wild videos. However, it lacks a benchmark for quantitatively comparing different HPE methods on a global level. To deal with this limitation, we propose SLOPER4D, the first large-scale urban-level human pose dataset with rich 2D/3D annotations.

3. SLOPER4D Dataset

SLOPER4D collects scene-aware 4D human data with our body-worn capturing system in urban scenes. In this section, we first introduce the data acquisition in Sec. 3.1, second, we detail the data construction and annotation process in Sec. 3.2, then we introduce the global optimization-based Sec. 3.3 method to obtain high quality both 3D/2D data, finally, we compare our dataset in Sec. 3.4 with the existing datasets and highlight our novelty.

3.1. Data Acquisition

Hardware setup. As shown in Fig. 1, during the data collection procedure, the scanning person follows the performer (IMUs wearer) and scans him with a LiDAR and a camera on the helmet. Additionally, Fig. 2 shows the hardware details of our capturing system. Regarding the sensor module, the 128-beams Ouster-os1 LiDAR and the DJI-Action2 camera are rigidly installed on the helmet. To capture raw human motions, we use Noitom's inertial MoCap product, PN Studio, to attach 17 wireless IMUs to the IMU wearer's body limbs, torso, and head. The camera's field of view (FOV) is 116°×84° and the LiDAR's FOV is 360°×45°. To make the performer within the LiDAR's FOV

Figure 2. **Our capturing system's hardware details.** The sensor module includes a LiDAR, a camera, and 17 body-attached IMU sensors. The storage module consists of a NUC11, a receiver, and a battery in the backpack.

as much as possible, we tilt the LiDAR down around 45°. Regarding the storage module, the scanning person's backpack places a wireless IMU data receiver, a 24V battery, and an Intel NUC11. The mini-computer NUC11 stores IMU data from the wireless receiver and point clouds from LiDAR in real-time. Videos are stored locally in the camera. The LiDAR and NUC11 are both powered by the battery.

Coordinate systems. Let's define three coordinate systems: 1) IMU coordinate system $\{I\}$: the origin is at LiDAR wearer's spine base at the starting time, and the $X/Y/Z$ axis is pointing left/upward/forward of the human. 2) LiDAR Coordinate system $\{L\}$: the origin is at the center of the LiDAR, and the $X/Y/Z$ axis is pointing right/forward/upward of the LiDAR. 3) Global/World coordinate system $\{W\}$: the origin is on the floor of the LiDAR wearer's starting position, and the $X/Y/Z$ axis is pointing right/forward/upward of the LiDAR wearer.

Calibration. Following the setup in [53], we use a chessboard to calibrate the camera intrinsic K_{in} and introduces a terrestrial laser scanner (TLS) to obtain accurate camera extrinsic parameter, K_{ex}. Due to the LiDAR point cloud being too sparse, we manually choose the corresponding points both on the 2D image and the TLS map registered to the point cloud, and then we solve the perspective-n-point (PnP) problem to obtain K_{in}. For every 3D scene, the calibration R_{WL}, which transforms $\{L\}$ to $\{W\}$ is manually set to make the ground's z-axis upward and height to zero for the starting position. By using singular value decomposition, the calibration R_{WI}, which transforms $\{I\}$ to $\{W\}$, is calculated through the similarity between IMU trajectory and LiDAR trajectory on the XY plane.

Synchronization. The synchronization of data from multiple sensors in human subject data is achieved through peak detection. Before and after the capture, the subject is asked to perform jumps. Then the peak height time in IMU is automatically detected and the peak times in the LiDAR and camera data are manually identified. Finally, all modalities are aligned by the peaks and downsampled to match the LiDAR frame rate of 20 Hz.

3.2. Data Processing

2D pose detection. We use Detectron [56] to detect and Deepsort [55] to track humans in videos. However, the tracking often fails due to the IMUs wearer entering/exiting the field of view or occlusions. To solve this problem, we manually assign the same ID for the tracked person in a video sequence. As for 3D point cloud reference, we project them on images according to the K_{ex}. However, due to the jitter brought by dynamic motions, the camera and the LiDAR are not perfectly rigidly connected. Thus, K_{ex} will be further optimized in Sec. 3.3.

LiDAR-inertial localization and mapping. The LiDAR-only method often fails in mapping because of the dynamic head rotation and crowded urban environments. Incorporating an IMU can compensate for motion distortion in a LiDAR scan p^L and provide an accurate initial pose. Using a LiDAR with an integrated IMU, and by combining Kalman filter-based lidar-inertial odometry [58] with factor graph-based loop closure optimization [7] [22], we successfully estimate the ego-motion of LiDAR and build the global consistency 3D scene map with n frame point clouds $P^L_{1:n} = \{p^L_1, \dots, p^L_n\}$. To provide accurate scene constrain in Sec. 3.3, we utilize the VDB-Fusion [47] to generate a clean scene mesh $\boldsymbol{S}$ that excludes moving objects.

IMUs pose estimation. We use SMPL [32] to represent the human body motion $M^I = \Phi(\theta^I, t^I, \beta) \in \mathbb{R}^{6890\times 3}$ in IMU coordinate space$\{I\}$, where pose parameter $\Theta^I_{1:n} = \{\theta^I_1, \dots, \theta^I_n\} \in \mathbb{R}^{72\times n}$ is composed of pelvis joint's orientation $R^I_{1:n} = \{r^I_1, \dots, r^I_n\} \in \mathbb{R}^{3\times n}$ and the other 23 joints' rotation relative to their parent joint. The $T^I_{1:n} = \{t^I_1, \dots, r^I_n\} \in \mathbb{R}^{3\times n}$ is the pelvis joint's translation and $\beta \in \mathbb{R}^{10}$ is a constant value representing a person's body shape. T and Θ are estimated by the commercial MoCap product, while β is obtained by using IPNet [3] to fit the scanned model captured by an iPhone13 Promax. Since the IMU are accurate locally but drift globally, T^I is used for raw calibration of the $\{I\}$ to $\{W\}$, and the initial global motion $M = M^W = R^{WI}M^I$ will be further optimized.

3.3. Data Optimization

To obtain precise and scene-plausible human motion M in the world coordinate system, we use scene geometry $\boldsymbol{S}$ with several physic-based terms to perform joint optimizations to find the optimal motion M^* that minimize $\mathcal{L}$. In a k-frame segment, the optimization is written as:

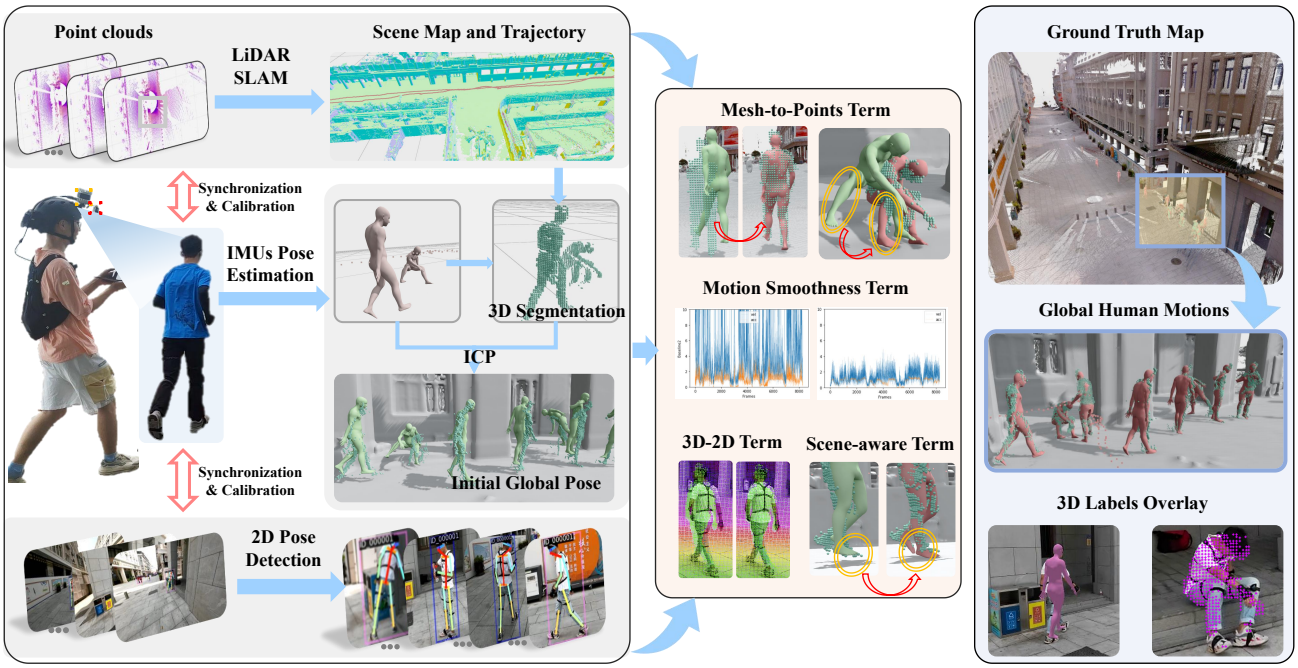

Figure 3. **The pipeline of the dataset construction.** The capturing system simultaneously collects multimodal data, including LiDAR, camera, and IMU data. Then they are further processed. A joint optimization approach with multiple loss terms is then employed to optimize motion locally and globally. As a result, we obtain rich 2D/3D annotations with accurate global motion and scene information.

$$
\begin{aligned}
M^*_{1:k} &= \arg\min_{M_{1:k}} \mathcal{L}(M_{1:k}, \boldsymbol{S}), \\
\mathcal{L} &= \mathcal{L}_{smt} + \lambda_{sc}\mathcal{L}_{sc} + \lambda_{pri}\mathcal{L}_{pri} + \lambda_{m2p}\mathcal{L}_{m2p}, \\
\mathcal{L}_{smt} &= \lambda_{trans}\mathcal{L}_{trans} + \lambda_{orit}\mathcal{L}_{orit} + \lambda_{jts}\mathcal{L}_{jts},
\end{aligned} \tag{1}
$$

where $\mathcal{L}_{smt}$ is a smoothness term, which consists of a translation loss $\mathcal{L}_{trans}$, an orientation loss $\mathcal{L}_{orit}$, and a joints loss $\mathcal{L}_{jts}$. $\mathcal{L}_{sc}$ is a scene-aware-contact term, $\mathcal{L}_{pri}$ is a pose prior term, and $\mathcal{L}_{m2p}$ is a mesh-to-points term. The λ_{sc}, λ_{pri}, λ_{trans}, λ_{orit}, λ_{jts}, and λ_{m2p} are loss terms' coefficients. $\mathcal{L}$ is minimized with a gradient descent algorithm.

Smoothness term. The objective of this term is to minimize the acceleration of the pelvis joint, the accelerations of the other 23 pelvis-relative joints, which is denoted as $J_{1:k} = \{\jmath_1, \ldots, \jmath_k\} \in \mathbb{R}^{69\times k}$, and the angular velocity of all joints to smooth human movements.

Scene-aware contact term. we compare the movement of every foot vertices in IMU motions M^I_k and label the foot as stable if its velocity is less than 0.1 m/s^2. Finally, the Chamfer Distance (CD) between this foot and its closest surface is expressed as the scene contact loss $\mathcal{L}_{sc}$.

Pose prior term. The poses estimated by IMUs are roughly accurate but will likely cause some misalignments to the end of the body limb due to the accumulating error. Hence, $\mathcal{L}_{prior}$ is used to constrain the Θ close to the initial value at the beginning of the optimization.

Mesh-to-points term. The point cloud p^L from the moving LiDAR provides strong prior depth information. However, though the SMPL mesh is watertight and complete, the human points are sparse and partial, which makes the registration methods such as ICP, not ideal as expected. To address this issue, we propose a viewpoint-based mesh-to-point loss function $\mathcal{L}_{m2p}$. First, we remove the hidden SMPL mesh faces from the LiDAR's viewpoint. Then we sample points, denoted as $P'_{1:k} = \{p'_1, \ldots, p'_k\}$, from the remaining faces by LiDAR resolution. The loss is defined as the Chamfer Distance from $P'_{1:k}$ to $P_{1:k}$.

All loss terms functions are detailed as follows:

$$
\begin{aligned}
\mathcal{L}_{trans} &= \frac{1}{k-2}\sum_{i=1}^{k-2} \|t_{i+2} - 2t_{i+1} + t_i\|_2^2, \\
\mathcal{L}_{jts} &= \frac{1}{k-2}\sum_{i=1}^{k-2} \|\jmath_{i+2} - 2\jmath_{i+1} + \jmath_i\|_2^2, \\
\mathcal{L}_{orit} &= \frac{1}{k-1}\sum_{i=1}^{k-1} \|r_{i+1} - r_i\|_2^2, \\
\mathcal{L}_{pri} &= \frac{1}{k}\sum_{i=1}^{k} \|\theta_i - R^{WI}\theta_i^I\|_2^2, \\
\mathcal{L}_{m2p} &= \frac{1}{k}\sum_{i=1}^{k} \sum_{\hat{p'}\in p'_i} \frac{1}{|p'_i|} \min_{\hat{p}\in p_i} \|\hat{p} - \hat{p'}\|_2^2.
\end{aligned} \tag{2}
$$

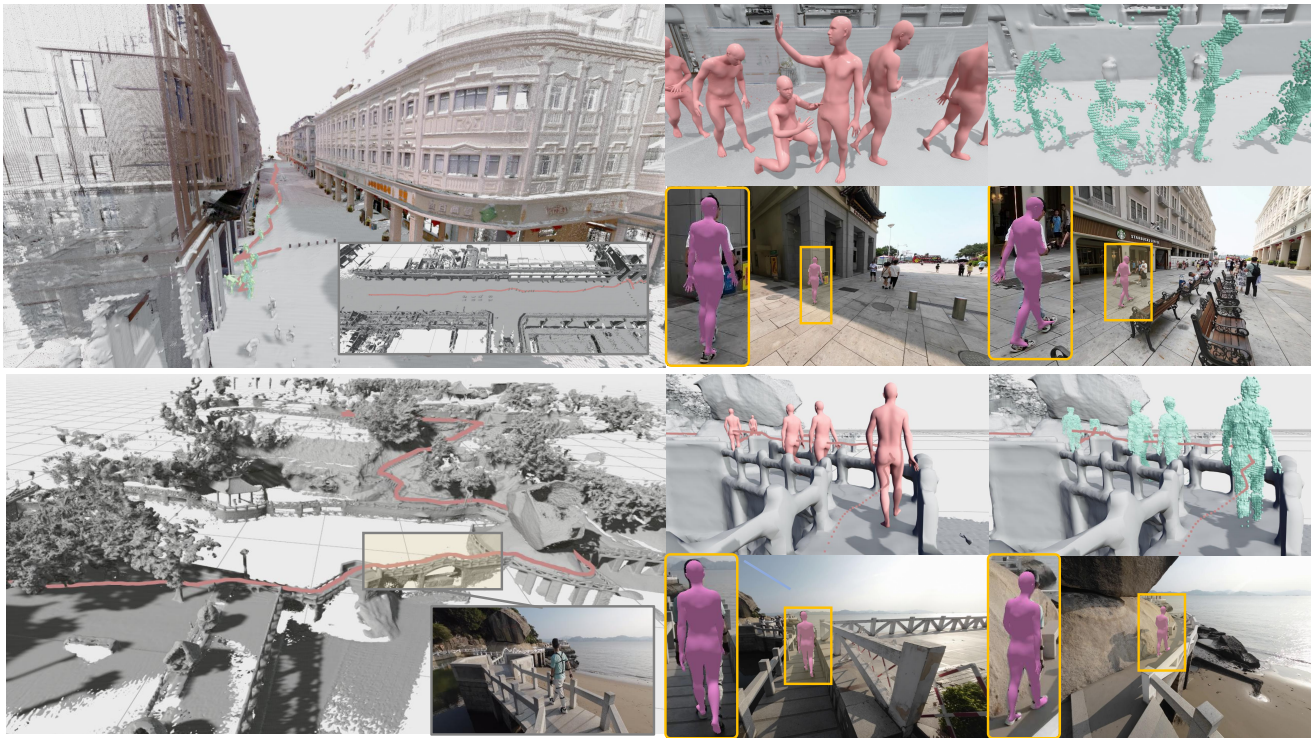

Figure 4. The diverse scenes and activities of our dataset. The images in the left column are our reconstructed scenes with human trajectories overlaid on them. The right images are the SMPL meshes overlaid on images / point clouds / scenes.

Camera extrinsic optimization. We aim to optimize extrinsic parameters K_{ex} for every frame by minimizing the $\mathcal{L}_{cam}$, which comprises of the keypoints loss $\mathcal{L}_{kpt}$ and the bounding box loss $\mathcal{L}_{box}$. The $\mathcal{L}_{kpt}$ measures the mean square error between the 2D human keypoints kpt^{2d} in the image and the 3D human keypoints kpt^{3d} of the optimized SMPL model projected to the image with K_{ex}; the $\mathcal{L}_{box}$ computes the Intersection over Union loss between the 2D human bounding box box^{2d} in the image and the 3D human bounding box box^{3d} projected to the image with K_{ex}.

$$\begin{aligned} K_{ex}^{*} &= \arg \min_{K_{ex} \in SE(3)} \mathcal{L}_{cam}(K_{ex}), \\ \mathcal{L}_{cam} &= \lambda_{kpt}\mathcal{L}_{kpt}(kpt^{3d}, kpt^{2d}, K_{ex}) + \\ &\quad \lambda_{box}\mathcal{L}_{box}(box^{3d}, box^{2d}, K_{ex}), \end{aligned} \tag{3}$$

where λ_{kpt} and λ_{box} are constant coefficients.

3.4. Dataset Comparison

SLOPER4D is the first large-scale urban-level human pose dataset with multi-modal capture data and rich human-scene annotations for GHPE. The head-mounted LiDAR and camera are utilized to simultaneously record the IMU-wearer's activities, including running outside, playing football, visiting, reading, climbing/descending stairs, discussing, borrowing a book, greeting, etc. The dataset consists of 15 sequences from 12 human subjects in 10 locations. There are a total of 100k LiDAR frames, 300k video frames, and 500k IMU-based motion frames captured over a total distance of more than 8 km and an area of up to 30,000 m^2. The results of our dataset are shown in Fig. 4.

For the captured person, we provide the segmentation of 3D points from LiDAR frames and 2D bounding boxes from images synchronized with LiDAR. We also provide 3D pose annotations with SMPL format. Compared to other datasets Tab. 1, it is worth mentioning that SLOPER4D provides the 3D scene reconstructions and accurate global translation annotations, allowing us to quantitatively study the scene-aware global pose estimation from both LiDAR and monocular videos. In addition to the dense 3D point cloud map reconstructed from the LiDAR, SLOPER4D provides the high-precision colorful point cloud map from a Terrestrial Laser Scanner (Trimble TX5) for better visualization and map comparison.

4. Experiments

In this section, we first evaluate SLOPER4D Dataset qualitatively, indicating that our dataset is solid enough to benchmark new tasks. Then we perform a cross-dataset evaluation to further assess our dataset's novelty on two tasks: LiDAR-based 3D HPE and camera-based 3D HPE. Finally, we introduce the new benchmark, GHPE, and perform experiments on GLAMR. More quantitative evaluations and experiments are in the supplementary material.

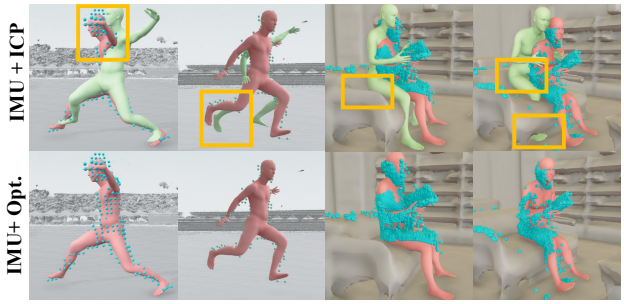

Figure 5. Comparison between our optimization results (red SMPL) and the ICP results (green SMPL). It shows the red SMPL aligns better with the cyan human points than the green SMPL.

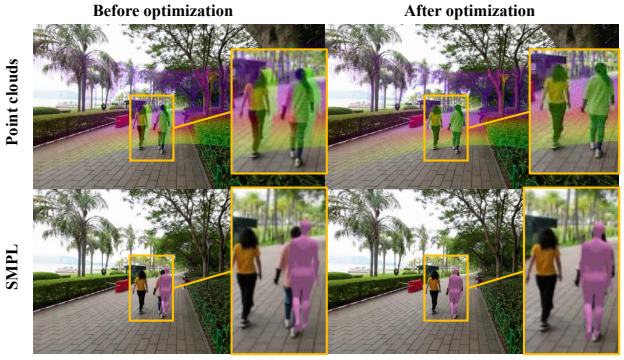

Figure 6. The comparison before (**left**) and after (**right**) extrinsic optimization by projecting the point clouds (upper) and SMPL (lower) are projected onto the image.

Training/Test splits. We split our data into training and test sets for LiDAR/Camera-based pose estimation. The training set of SLOPER4D contains eleven sequences of data with a total of 80k LiDAR frames and corresponding RGB frames. The test set has four sequences of data with around 20k LiDAR frames and corresponding RGB frames. For global pose estimation, we select three challenging scenarios for evaluation. The first one is a single-person football training scenario with highly dynamic motions. The second one is running along a coastal runway. The third one is a garden tour involving daily motions.

Evaluation metrics. For 3D HPE, we employ Mean per joint position error (MPJPE) and Procrustes-aligned MPJPE (PA-MPJPE) for evaluation. MPJPE is the mean euclidean distance between the ground-truth and predicted joints. PA-MPJPE first aligns the predicted joints to the ground-truth joints by carrying out rigid transformation based on Procrustes analysis and then calculates MPJPE. For global trajectory evaluation, we utilize Absolute Trajectory Error (ATE) and the Relative Pose (the pose refers to orientation here) Error (RPE) in visual SLAM systems [11], where the ATE is well-suited for measuring the global localization and, in contrast, the RPE is suitable for measuring the system's drift, for example, the drift per second. Global MPJPE (G-MPJPE) is MPJPE calculated by placing the SMPL model in the global coordinates.

Qualitative evaluation. For the human pose qualitative evaluation, we project the SMPL to the image and visualize the 3D human with corresponding LiDAR points in 3D space (shown in Fig. 4). The results demonstrate that the 3D human mesh aligns well with 3D environments and 2D images. As a large-scale urban-level human pose dataset, SLOPER4D provides multi-modal capture data and rich human-scene annotations, as well as diverse challenging human activities in large scenes. To evaluate our optimization method, we first compare our method with the results from ICP. As shown in Fig. 5, the scene-aware constraints and human mesh-to-points constraint efficiently optimize the local poses, global translation, and even the orientation error from IMU. To show the effectiveness of the camera extrinsic optimization, we report the results in Fig. 6. The 2D projecting error was visually lowered after optimization.

4.1. Cross-Dataset Evaluation

Train \ Test	LH26M		Ours	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
LH26M	**79.3**	**67.0**	228.7	149.9
Ours	212.3	128.3	86.1	65.1
LH26M + Ours	85.5	72.0	**79.2**	**60.1**

(a) LiDAR-based 3D pose estimation with LiDARCap [27].

Train \ Test	3DPW		Ours	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
VIBE [24]	93.5	56.5	102.5	66.2
Ours + AMASS	124.3	66.8	**86.6**	**52.4**
w. 3DPW	**83.0**	**52.0**	90.0	58.3
HbryIK [26]	88.7	49.3	104.9	57.0
w. Ours	87.3	49.2	67.6	44.2
w. 3DPW	**71.3**	**41.8**	75.8	50.0
w. 3DPW + Ours	76.4	46.7	**66.2**	**42.8**

(b) Camera-based 3D pose estimation.

Table 2. Cross-dataset evaluation results with different modalities. The LH26M in (a) refers to LiDARHuman26M dataset from LiDARcap. VIBE is pre-trained on AMASS [33], MPI-INF-3DHP [35] InstaVariety [19], PoseTrack [1], PennAction [65]. HbryIK is pre-trained on H36M, MPI-INF-3DHP, MSCOCO [29]

We evaluate root-relative 3D human pose estimation with different modalities, namely the LiDAR and the camera. 3DPW is an in-the-wild human motion dataset that is most related to us. With VIBE, we cross-evaluated our dataset's camera modal by using 3DPW. LiDARHuman26M is a lidar-based dataset for long-range human pose estimation. We can cross-evaluate our dataset's LiDAR modal with it. Tab. 2(a) shows the evaluation results on LiDAR-based 3D pose estimation task and Tab. 2(b) shows the results on camera-based 3D pose estimation. Taking the results from Tab. 2(a), for example, when the model is trained from another dataset only, the errors are the largest. But the error will be further reduced by around 60% when training on LiDARHuman26M and our dataset together. It

suggests a domain gap exists between different LiDAR sensors, and both datasets complement each other. The results of another task show that the pre-trained VIBE model generalizes better on 3DPW than on our dataset. But the error on 3DPW increases when finetuned on our dataset, while the error decreases on our dataset. This suggests that the pre-trained model complements SLOPER4D better than the opposite. Comparing the results across different modalities, the error on our dataset from the method trained on mixed LiDAR point cloud datasets is 13% lower than the method trained on the images.

4.2. Benchmark on Global Human Pose Estimation

Sequence	Metric	RMSE ↓	*mean*	*std.*	*max*
Football	ATE	3.26	2.85	1.58	11.83
Running001	ATE	29.48	25.55	14.72	56.07
Garden001	ATE	**2.86**	2.57	1.26	6.55
Football	RPE	0.08	0.06	0.05	1.34
Running001	RPE	0.40	0.35	0.19	1.04
Garden001	RPE	**0.06**	0.04	0.04	0.71

Table 3. Global trajectory evaluation of GLAMR. Unit: m.

Sequence	Scale	MPJPE ↓	PA-MPJPE ↓	G-MPJPE ↓
Football	11.83	264.6	118.5	5268.7
Running001	56.07	652.1	119.6	32329.3
Garden001	6.55	**139.4**	**86.3**	**4407.0**

Table 4. GHPE results from GLAMR. Unit: mm.

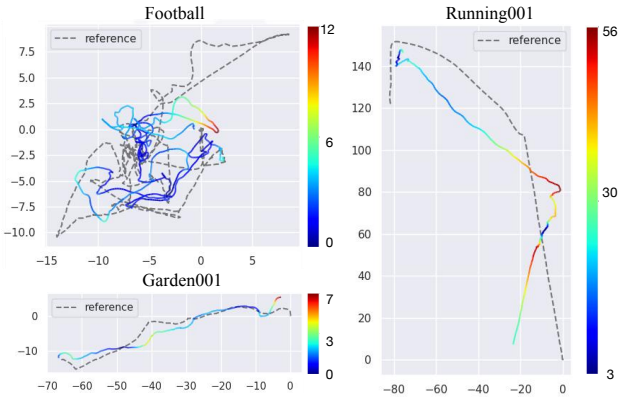

Figure 7. The ATE error (m) mapped on the GT trajectory. The color represents the error according to the color bar.

In this subsection, we benchmark the GHPE task of GLAMR [61] on SLOPER4D. GLAMR is a global occlusion-aware method for 3D global human mesh recovery from dynamic monocular cameras. For the scale uncertainty of the monocular camera, we compute the affine matrix from the estimated trajectory to the ground truth trajectory and rotate, translate and scale the estimated trajectory before error computation.

Tab. 3 reports the global trajectory error with ATE and RPE, Tab. 4 reports the global human pose metric, and Fig. 7 shows the ATE error mapped on GT trajectory. Comparing the results on the three scenes, the *football* and *Garden001* have a significantly lower RPE in the global scene. In comparison, GLAMR performs the worst on the running scene, with an ATE's RMSE of 29.48 m. This scene has the largest area size and the highest human pace. GLAMR achieves a low PA-MPJPE of 86.3mm on *Garden001*, a sequence with daily walking and visiting motions. It's the first time that we have tested the GPHE on such large outdoor scenes. GLAMR achieves relatively better results on daily human motion while performing worse on high-dynamic activities in the wild. The interesting point is that the trajectory tendency is pretty similar to the reference, even in dynamic football training motions, which demonstrates the ability of GLAMR to be a baseline. It is expected that more research will focus on GHPE in real-world interactive scenarios, and the experiments show our SLOPER4D's potential to promote urban-level GHPE research.

5. Discussions

Limitations. Firstly, SLOPER4D is limited to single-person capture though it perceives multiple-person data. Secondly, the camera and LiDAR are not synchronized online, causing tedious offline work if the camera loses frames even with a low time offset ($<$50 ms). Finally, texture information from the camera is not fully exploited for color and texture reconstruction of scenes and humans. In our future work, we will propose an online synchronization algorithm and extend our work to multiple-person capturing.

Conclusions. We propose the first large-scale urban-level human pose dataset with multi-modal capture data and rich human-scene annotations. Based on our proposed new dataset, we benchmark two critical tasks, camera-based 3D HPE and LiDAR-based 3D HPE. SLOPER4D also benchmarks the GHPE task. The results demonstrate the potential of SLOPER4D in boosting the development of these areas. Our work contributes to extending motion capture to large global scenes based on the current methods and datasets. We hope this work will foster future creation and interaction in urban environments.

Acknowledgements. This work was supported in part by the Natural Science Foundation of China (No.62171393), the Fundamental Research Funds for the Central Universities (No.20720220064), FuXiaQuan National Independent Innovation Demonstration Zone Collaborative Innovation Platform (No.3502ZCQXT2021003), and NSFC (No.62206173). We thank Zhiyong Wang for helping us incorporate FAST-LIO2 into our mapping system. We also acknowledge support from Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 7

[2] Sébastien Bauwens, Harm Bartholomeus, Kim Calders, and Philippe Lejeune. Forest inventory with terrestrial lidar: A comparison of static and hand-held mobile laser scanning. *Forests*, 7(6):127, 2016. 3

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, aug 2020. 4

[4] Michael Bosse, Robert Zlot, and Paul Flick. Zebedee: Design of a spring-mounted 3-d range sensor with application to mobile mapping. *IEEE Transactions on Robotics*, 28(5):1104–1119, 2012. 3

[5] Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6792–6802, June 2022. 2, 3

[6] Hanieh Deilamsalehy and Timothy C Havens. Sensor fused three-dimensional localization using imu, camera and lidar. In *2016 IEEE SENSORS*, pages 1–3. IEEE, 2016. 3

[7] Frank Dellaert and GTSAM Contributors. borglab/gtsam. https://github.com/borglab/gtsam, May 2022. 4

[8] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016. 3

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[10] Patrick Geneva, Kevin Eckenhoff, Yulin Yang, and Guoquan Huang. Lips: Lidar-inertial 3d plane slam. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 123–130. IEEE, 2018. 3

[11] Michael Grupp. evo: Python package for the evaluation of odometry and slam. https://github.com/MichaelGrupp/evo, 2017. 7

[12] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021. 2, 3

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

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022. 2, 3

[15] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37(6):185:1–185:15, nov 2018. 3

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2, 3

[17] Tomoya Kaichi, Tsubasa Maruyama, Mitsunori Tada, and Hideo Saito. Resolving position ambiguity of imu-based human pose with a single rgb camera. *Sensors*, 20(19):5453, 2020. 2

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

[19] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[20] Samer Karam, George Vosselman, Michael Peter, Siavash Hosseinyalamdary, and Ville V. Lehtola. Design, calibration, and evaluation of a backpack indoor mobile mapping system. *Remote. Sens.*, 11:905, 2019. 3

[21] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 3

[22] Giseop Kim and Ayoung Kim. Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4802–4809, 2018. 4

[23] Giseop Kim, Yeong Sang Park, Younghun Cho, Jinyong Jeong, and Ayoung Kim. Mulran: Multimodal range dataset for urban place recognition. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Paris, May 2020. 3

[24] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 3, 7

[25] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 3

[26] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse

kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 3, 7

[27] Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20502–20512, 2022. 2, 3, 7

[28] Jiarong Lin and Fu Zhang. Loam livox: A fast, robust, high-precision lidar odometry and mapping package for lidars of small fov. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3126–3131. IEEE, 2020. 3

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[30] 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. *international conference on 3d vision*, 2020. 3

[31] Timothy Liu, Matthew Carlberg, George Chen, Jacky Chen, John Kua, and Avideh Zakhor. Indoor localization and visualization using a human-operated backpack system. *2010 International Conference on Indoor Positioning and Indoor Navigation*, pages 1–10, 2010. 3

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 2015. 2, 4

[33] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2, 7

[34] Charles Malleson, Marco Volino, Andrew Gilbert, Matthew Trumble, John Collomosse, and Adrian Hilton. Real-time full-body motion capture from video and imus. In *2017 Fifth International Conference on 3D Vision (3DV)*, 2017. 3

[35] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 7

[36] Roberto Opromolla, Giancarmine Fasano, Giancarlo Rufino, Michele Grassi, and Al Savvaris. Lidar-inertial integration for uav localization and mapping in complex environments. In *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 649–656. IEEE, 2016. 3

[37] Noha Radwan, Abhinav Valada, and Wolfram Burgard. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. *IEEE Robotics and Automation Letters*, 3(4):4407–4414, 2018. 3

[38] Daniel Roetenberg, Henk Luinge, and Per Slycke. Moven: Full 6dof human motion tracking using miniature inertial sensors. *Xsen Technologies, December*, 2(3):8, 2007. 3

[39] Daniel Roetenberg, Henk Luinge, and Per Slycke. Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. *Xsens Motion Technologies BV, Tech. Rep*, 1, 2009. 2, 3

[40] Youngwoo Seo and Chih-Chung Chou. A tight coupling of vision-lidar measurements for an effective odometry. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1118–1123. IEEE, 2019. 3

[41] Tixiao Shan and Brendan Englot. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4758–4765. IEEE, 2018. 3

[42] Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Daniela Rus. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5135–5142. IEEE, 2020. 3

[43] Tixiao Shan, Brendan Englot, Carlo Ratti, and Daniela Rus. Lvi-sam: Tightly-coupled lidar-visual-inertial odometry via smoothing and mapping. *arXiv preprint arXiv:2104.10831*, 2021. 3

[44] Young-Sik Shin, Yeong Sang Park, and Ayoung Kim. Dvl-slam: sparse depth enhanced direct visual-lidar slam. *Autonomous Robots*, 44(2):115–130, 2020. 3

[45] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010. 2

[46] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13, 2017. 2

[47] Ignacio Vizzo, Tiziano Guadagnino, Jens Behley, and Cyrill Stachniss. Vdbfusion: Flexible and efficient tsdf integration of range sensor data. *Sensors*, 22(3), 2022. 4

[48] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. Practical motion capture in everyday surroundings. *ACM Transactions on Graphics (TOG)*, 26(3):35, 2007. 3

[49] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2, 3

[50] Timo von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and imus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1533–1547, 2016. 3

[51] Timo von Marcard, Bodo Rosenhahn, Michael Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference*

of the European Association for Computer Graphics (Eurographics), pages 349–360, 2017. 3

[52] Bing Wang, Changhao Chen, Chris Xiaoxuan Lu, Peijun Zhao, Niki Trigoni, and Andrew Markham. Atloc: Attention guided camera localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10393–10401, 2020. 3

[53] Chenglu Wen, Yudi Dai, Yan Xia, Yuhan Lian, Jinbin Tan, Cheng Wang, and Jonathan Li. Toward efficient 3-d colored mapping in gps-/gnss-denied environments. *IEEE Geoscience and Remote Sensing Letters*, 17(1):147–151, 2019. 4

[54] Chenglu Wen, Yudi Dai, Yan Xia, Yuhan Lian, Jinbin Tan, Cheng Wang, and Jonathan Li. Toward efficient 3-d colored mapping in gps-/gnss-denied environments. *IEEE Geoscience and Remote Sensing Letters*, 17(1):147–151, 2020. 3

[55] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 4

[56] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 4

[57] Lan Xu, Yebin Liu, Wei Cheng, Kaiwen Guo, Guyue Zhou, Qionghai Dai, and Lu Fang. Flycap: Markerless motion capture using multiple autonomous flying cameras. *IEEE transactions on visualization and computer graphics*, 24(8):2284–2297, 2017. 2

[58] Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38:2053–2073, 2022. 4

[59] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[60] Hongwei Yi, Chun-Hao P Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J Black. Human-aware object placement for visual environment reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3959–3970, 2022. 2, 3

[61] 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. 3, 8

[62] Ji Zhang and Sanjiv Singh. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and Systems*, volume 2, 2014. 3

[63] Ji Zhang and Sanjiv Singh. Visual-lidar odometry and mapping: Low-drift, robust, and fast. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2174–2181. IEEE, 2015. 3

[64] 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)*, Oct. 2022. 2, 3

[65] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 2248–2255, 2013. 7

[66] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *European Conference on Computer Vision*, pages 550–566. Springer, 2020. 3

[67] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Quionghai Dai, Lu Fang, and Yebin Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[68] Xingxing Zuo, Patrick Geneva, Woosik Lee, Yong Liu, and Guoquan Huang. Lic-fusion: Lidar-inertial-camera odometry. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5848–5854. IEEE, 2019. 3