HSC4D: Human-centered 4D Scene Capture in Large-scale Indoor-outdoor Space Using Wearable IMUs and LiDAR

Yudi Dai[1]   Yitai Lin[1]   Chenglu Wen[1,*]   Siqi Shen[1]   Lan Xu[1]   Jingyi Yu[1]   Yuexin Ma[1]   Cheng Wang[1]

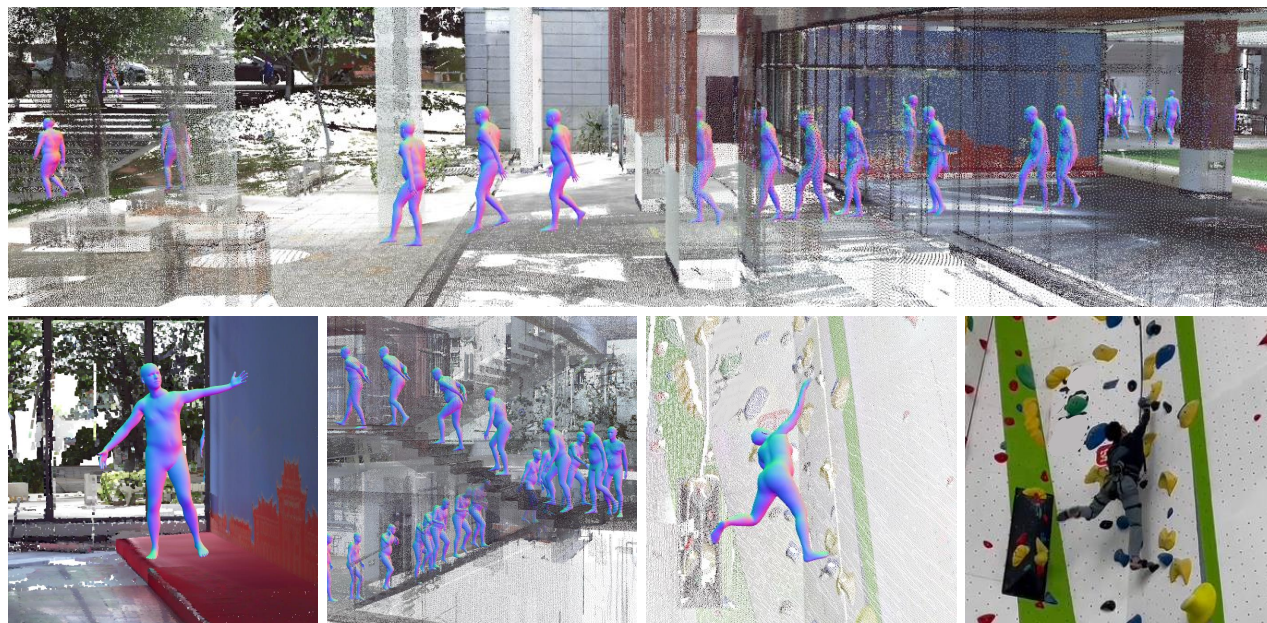[1]Xiamen University, China   [2]ShanghaiTech University, China

Figure 1. 3D human motions with accurate localization in diverse, challenging scenes. The top line shows a person walking from indoor to outdoor. The bottom row figures show the challenge cases, where the bottom right figure is the rock climbing's third-view from the camera.

Abstract

We propose Human-centered 4D Scene Capture (HSC4D) to accurately and efficiently create a dynamic digital world, containing large-scale indoor-outdoor scenes, diverse human motions, and rich interactions between humans and environments. Using only body-mounted IMUs and LiDAR, HSC4D is space-free without any external devices' constraints and map-free without pre-built maps. Considering that IMUs can capture human poses but always drift for long-period use, while LiDAR is stable for global localization but rough for local positions and orientations, HSC4D makes both sensors complement each other by a joint optimization and achieves promising results for long-term capture. Relationships between humans and environments are also explored to make their interaction more realistic. To facilitate many down-stream tasks, like AR, VR, robots, autonomous driving, etc., we propose a dataset containing three large scenes (1k-5k m^2) with accurate dynamic human motions and locations. Diverse scenarios (climbing gym, multi-story building, slope, etc.) and challenging human activities (exercising, walking up/down stairs, climbing, etc.) demonstrate the effectiveness and the generalization ability of HSC4D. The dataset and code is available at lidarhumanmotion.net/hsc4d.

1. Introduction

The development of digital society is overwhelming because it can enrich peoples' life by serving Augmented Reality, Virtual Reality, smart city, robots, autonomous driving, etc. Humans and environments are two main components for creating the digital world. Current research works always separate dynamic human motions and static environments. Actually, taking account their interactions can help improve both capture accuracy. It is a trend to directly capture the whole scene with consecutive human activities.

To capture human motions, IMU sensors are widely used and always be mounted on different parts of the human body, like arms, legs, feet, head, etc. It can capture accu-

---

*Corresponding author.

rate short-term motions but suffer from severe drift with the acquisition time increasing. Some methods [6,13,39,47,48] utilize extra external RGB or RGBD cameras as a remedy to improve the accuracy, but result in limited capture space, human activities, and interactions. HPS [9] uses a head-mounted camera, which looks outwards like the human eyes, to complement IMUs in global localization. Without the constraint of external cameras, it can recover the full-body pose and register the human in large 3D scan of real scenes. However, HPS requires pre-built maps and a huge image database for self-localization.

For accurate localization and mapping [51], LiDAR is the most applicable sensor in current days, which is popular for mobile robots and autonomous vehicles. LiDAR is also extensively used for large-scale scene scans. Although there are many LiDAR-captured datasets, including indoor scenes [4, 30] and large-scale outdoor scenes [7, 23], they focus on scene understanding and 3D perception, ignoring accurate human poses. PedX [16] provides 3D poses of pedestrians by using SMPL [22] parameterization for joint locations of instances on third-person-view images, which is not accurate as IMUs. Furthermore, it focuses on traffic scenes and is not applicable for generating diverse 3D human motions.

Taking advantage of IMUs-based motion caption and LiDAR-based localization and scene capture, we propose Human-centered 4D Scene Capture (HSC4D) to accurately and efficiently create a dynamic digital world with consecutive human motions in indoor-outdoor scenes. Using only body-mounted sensors, HSC4D is space-free and pose-free, and the interaction between humans and the environment inside is also free, which makes it possible to capture most of the human-involved real-world scenes. Compared with camera-based localization, LiDAR is more precise for global localization, which dramatically reduces the drift of IMUs, and does not need pre-built maps. IMUs can improve the accuracy of LiDAR-captured local trajectories, where the error is caused by the jitter of the body. Making use of the complement of both sensors, we propose a joint optimization to improve the performance of motion estimation and human-scene mapping by considering several physical constraints.

To facilitate further research and down-stream applications, we propose a dataset containing three large scenes (1k-5k m^2) with accurate dynamic human motions and locations. As Fig. 1 shows, the dataset contains diverse scenarios, like climbing gym, multi-story building, slope, etc., and challenging human activities, such as exercising, walking up/down stairs, climbing, etc. Accurate human poses and natural interactions between human and the environment demonstrate the effectiveness and the generalization ability of HSC4D.

Our contributions are summarized as follows:

- Based on body-mounted IMUs and LiDAR, we propose Human-centered 4D Scene Capture (HSC4D) for creating a human-centered dynamic digital world, which is space-free, pose-free, and interaction-free.
- We propose a joint optimization method by integrating LiDAR SLAM results and IMU poses with scene constraints, resulting in natural human motions and accurate global localization in large scenes.
- We provide a new dataset containing LiDAR point cloud of large-scale scenes, IMU data of human poses, and the results of poses and mapping by our optimization, which also demonstrates the effectiveness and the generalization ability of HSC4D.

2. Related work

2.1. IMU sensors for human pose estimation

IMU sensors have been widely used to capture human motions [11, 29, 38, 41]. However, IMU-based methods suffer from severe drift over time. To improve the pose estimation accuracy, some methods [6, 13, 24, 39, 40] utilize extra external RGB or RGBD cameras as a remedy. Helten *et al*. [10] combined two RGB-D cameras with IMUs to perform local pose optimization. Hybrid-Fusion [54] has achieved more accurate motion tracking performance by combining an RGBD camera with multiple IMUs. 3DPW [39] uses a single hand-held RGB camera and IMUs to optimize human pose for a certain period of frames simultaneously. Constraints from external cameras assist in recovering more accurate 3D poses but result in limited capture space, human activities, and interactions. HPS [9] uses a first-view head-mounted camera to self-localize the 3D pose from IMUs to the scene. However, HPS requires pre-built maps and an image database for self-localization. Instead, we use only body-mounted IMUs and LiDAR. Without any external devices' constraints and any pre-built maps, we achieve promising results for long-term human motion capture.

2.2. Human self-localization methods

Human self-localization aims at estimating the 6-DoF of the human subject with carrying devices. The received signal strength (RSS) fingerprinting-based methodologies [1, 2, 18] are widely used for indoor human localization. However, these methods need external receivers and are limited to the indoor space. Some image-based methods [15, 28, 42] regress locations directly from a single image with a pre-built map. Still, the scene-specific property makes them hard to generalize to unseen scenes. Some methods integrate IMU as an aid sensor [25, 34] to improve accuracy. With robustness and low drift, LiDAR-based localization has been successfully applied in indoor [27, 44] and outdoor [19, 37, 49, 50] scenes. To localize the human

subject, LiDAR are designed as backpacked [14, 21, 46] and hand-held [3]. LiDAR-based localization systems are usually big pieces of equipment and would affect human motion. We design a lightweight hip-mounted LiDAR to rigidly connect with the human body, achieving human self-localization in both large indoor and outdoor scenes.

2.3. LiDAR-based mapping methods

LiDAR is currently the most applicable sensor for 3D mapping. As a pioneer, zhang *et al.* [51] proposed LOAM, a real-time odometry and mapping method using a LiDAR, greatly boosting the 3D mapping research. Some methods [12, 20, 33, 43] further improve LOAM mapping for specific scenes and sensors. LeGo-LOAM [33] is a ground-optimized version, which requires to keep the LiDAR horizontal. LiDAR-based methods tend to fail when the Z-axis jitters severely. To address the drift problem and improve robustness, more sensors, such as visual sensors [32, 36, 52], IMU [8, 26, 34], or both [5, 35, 55], have been integrated in mapping task. To make the system lighter and able to work wirelessly, we propose a LiDAR-only method for localization and mapping in the scene. The joint optimization result with scene and IMU poses will further improve the LiDAR mapping result.

3. System setup

3.1. Notations and Task Description

The problem addressed in this paper is to estimate the 3D human motion with a 3D spinning LiDAR and IMUs in a large unknown scene and build a map for it, where human motion includes local 3D pose and global localization.

Notations. The N frames human motion is represented as $M = (T, \theta, \beta)$, where T is the $N \times 3$ translation parameter, θ is the $N \times 24 \times 3$ pose parameter, and β is the $N \times 10$ shape parameter. We assume that β is constant during a data recording. The 3D point cloud scene is represented as S. We use right subscript $k, k \in Z^+$ to indicate the index of a frame. We use the Skinned Multi-Person Linear (SMPL) body model [22] $\Phi(\cdot)$ to map M_k to human mesh models $V_k, V_k \in \mathbb{R}^{6890 \times 3}$.

Let us define three coordinate systems: 1) IMU coordinate system $\{I\}$: origin is at the hip joint of the first SMPL model, and $X/Y/Z$ axis is pointing to the right/upward/forward of the human. 2) LiDAR Coordinate system $\{L\}$: origin is at the center of the LiDAR, and $X/Y/Z$ axis is pointing to the right/forward/upward of the LiDAR. 3) Global coordinate system $\{W\}$: the first LiDAR frame's coordinate.

Task definition. Given a sequence of LiDAR sweep $P_k^L, k \in Z^+$ in $\{L\}$ and a sequence of 3D Human motion M_k^I in $\{I\}$, compute the human motion M_k^W in $\{W\}$ and build the 3D scene S with P_k^I.
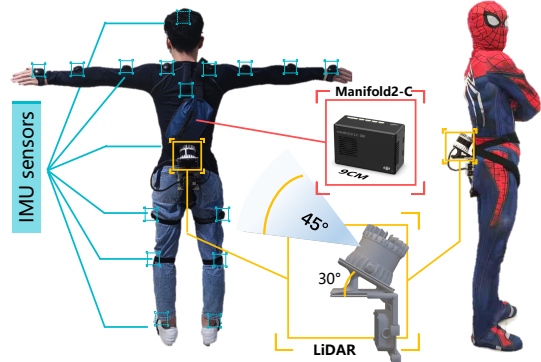
3.2. System design

Figure 2. Overview of our capturing system, which includes 17 body-attached IMU sensors, a LiDAR sensor, and a mini-computer.

Hardware. We use a 64-beams Ouster LiDAR to acquire 3D point clouds P_k^L, and Noitom's inertial MoCap product PN Studio to obtain human motion M_k^I. The PN Studio uses 17 IMUs attached to the body limbs and a wireless receiver to acquire data. To make the LiDAR work wirelessly, we connect the LiDAR to a $9cm \times 6cm \times 3cm$ DJI Manifold2-C mini-computer and use a 24V mobile power to charge the LiDAR and the computer. To ensure a lightweight and precise capturing system, We modified all cables and designed a $10cm \times 10cm$ L-shape bracket to mount the LiDAR package. The battery and Manifold2 are stored in a small bag on the human's back. The LiDAR is worn tightly close to the hip bone, making the origins of $\{I\}$ and $\{L\}$ as close as possible. Thus, we assume that LiDAR and IMUs have a rigid transformation.

The Ouster LiDAR has a 360° horizon view and a 45° vertical view. However, due to the occlusion caused by back and swing arms, the horizon field of view is reduced, ranging from 150° to 200°. To avoid most laser points hitting the nearby ground, we tilt up the LiDAR for 30° to get a good vertical scanning view.

4. Approach

We first obtain the 3D human motion output by the inertial MoCap system. Second, we estimate the ego-motion of LiDAR and build a 3D scene map S through point cloud data P_k^I. Then we perform a data initialization to prepare data for further optimization. Later we perform a graph-based optimization to fuse LiDAR trajectory and IMU trajectory. Finally, by combining the LiDAR data, IMUs data, and 3D scene, a joint optimization is performed to give the human motion M and an optimized scene.
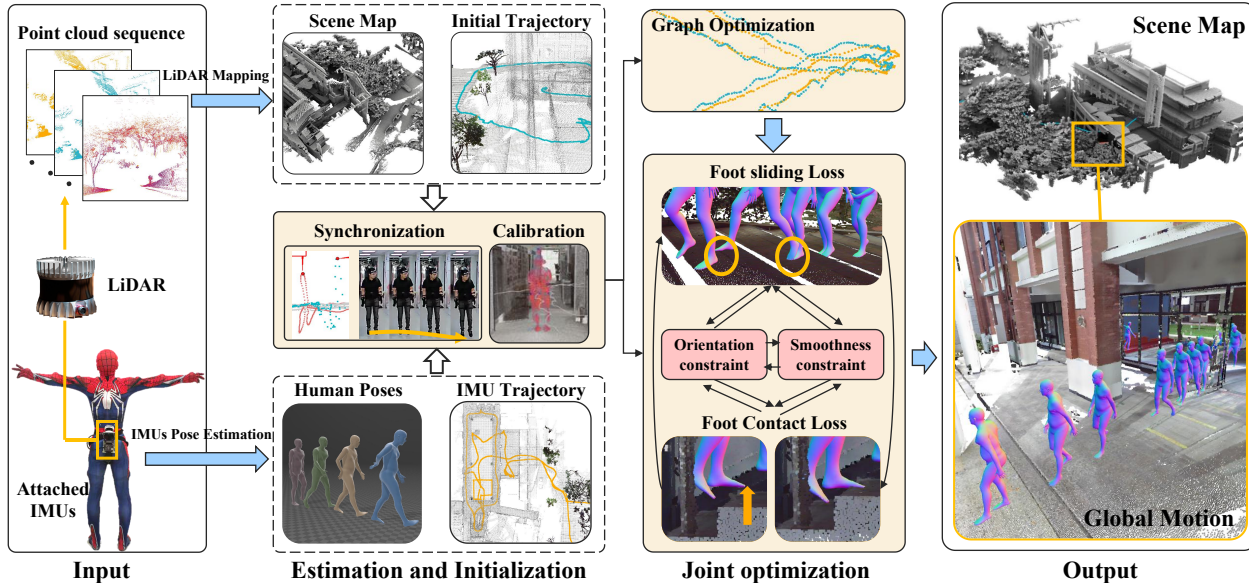
Figure 3. **Overview of our method.** The inputs are point cloud sequence and IMUs data. IMUs pose estimation and LiDAR mapping are performed, respectively. Synchronization and calibration are applied to prepare data for further optimization. Finally, graph-based optimization and joint optimization are performed to produce the global motion in the scene map.

4.1. IMUs Pose Estimation

This subsection aims to estimate the motion $M^I = (T^I, \theta^I, \beta)$ in IMU coordinate $\{I\}$, where T^I and θ^I are provided by the commercial MoCap product. Pose parameter θ^I is composed of the hip joint's orientation R^I relative to the start frame and other joints' rotation relative to their parent joint. T_k^I indicates the k-th frame translation relative to the start. Since IMU is accurate in a short period, the relative value of T^I and R^I can be used in the optimization.

4.2. LiDAR Localization and Mapping

Building a map using LiDAR data is challenging in this scene because the LiDAR jitters as the human walking and human body occludes the field of scanning view. By employing LiDAR-based SLAM methods [45,51], we estimate the ego-motion of LiDAR and build the 3D scene map S with $P_k^L, k \in Z^+$ in $\{L\}$. We first exact planer and edge feature points in every LiDAR scan P_k^L and keep updating the feature map. Similar to [45], we skip frame to frame odometry and only perform frame to map registration because the mapping process can run offline. Finally, Lidar's ego-motion T^W and R^W, and the scene map S are computed. The mapping function is denoted as:

$$T^W, R^W, S = \mathcal{F}(P_{1:N}^L) \qquad (1)$$

4.3. Optimization initialization

Coordinate calibration. To obtain the rigid offset from LiDAR to IMU and make all coordinate systems aligned, we perform following steps: First, the human stands as an A-pose before capture, and the human's face direction is regarded as scene's Y-axis direction. After capturing, we rotate the scene cloud Z-axis perpendicular to the starting position's ground. Last, we translate the scene to make its origin to the first SMPL model's origin on the ground. Li-DAR's ego motion T^W and R^W are translated and rotated as the scene does. To now, LiDAR data are calibrated to $\{W\}$. The pitch, roll, and translation of IMU are calibrated to the world coordinate. The IMU's yaw will be further refined in Sec. 4.4.

Time synchronization. Firstly, we ask the subject to jump at the starting place of every capture. Secondly, we automatically locate the peaks in both T^W and T^I based on their z value. Then, we can synchronize the LiDAR and IMU according to the two peaks' timestamps. Finally, we resample the IMU data (100Hz) to the same frame rate as LiDAR (20Hz).

4.4. Graph optimization data fusion

As seen from Fig. 4, IMUs drift severely over time and fail when the scene's height change, while the LiDAR localizes correctly but jitters at local movement. To estimate a more smooth and stable trajectory, we utilize both data's advantages. Our strategy is presented as follows: 1) first mark the frame in T^W that exceeds x (1.2~2.0) times of IMU velocity and the local fitted value as the outliers, 2) treat the remained (R^W, T^W) as landmarks, and then segment T^I every five seconds, 3) The nodes (IMU or landmark poses) and edges (a relative transformation between two nodes) construct a graph, 4) finally perform a graph op-
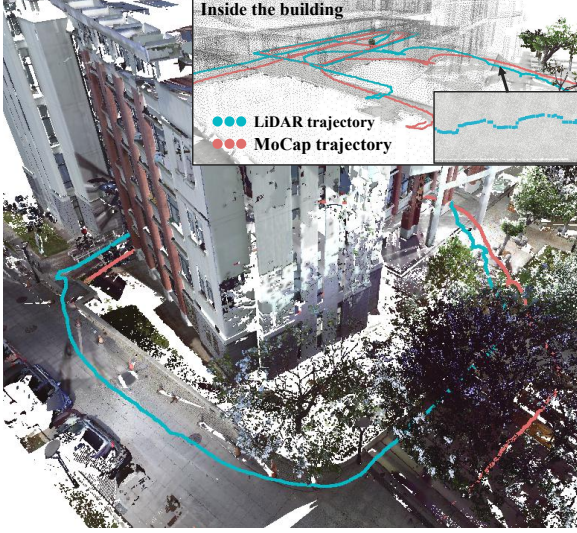
Figure 4. Comparison of trajectories of IMU and LiDAR.

timization [17] method to couple T^I and T^W.

4.5. Joint optimization

To obtain accurate and scene-natural human motion $\boldsymbol{M} = (T, \theta)$, and a higher quality scene cloud $\boldsymbol{S}$, we perform a joint optimization method by using scene $\boldsymbol{S}$ and physics constraints. Then we send T back to mapping function $\mathcal{F}$ as an initial value to create a new scene $\boldsymbol{S}_{opt}$. We use the following four constraints: the foot contact constraint $\mathcal{L}_{cont}$ encouraging the human standing on the ground, the sliding constraints $\mathcal{L}_{sld}$ eliminating the human walk sliding, the orientation constraint $\mathcal{L}_{ort}$ from R^I making the rotation smooth, and the smoothness constraint $\mathcal{L}_{smt}$ making the translation smooth. The optimization is expressed as:

$$\mathcal{L} = \lambda_{cont}\mathcal{L}_{cont} + \lambda_{sld}\mathcal{L}_{sld} + \lambda_{ort}\mathcal{L}_{ort} + \lambda_{smt}\mathcal{L}_{smt}$$
$$\boldsymbol{M} = arg\min_{\boldsymbol{M}}\mathcal{L}(\boldsymbol{M}|T^I, \theta^I, R^W, \boldsymbol{S}) \quad (2)$$
$$\boldsymbol{S}_{opt} = \mathcal{F}(P^L_{1:N}, \boldsymbol{M}),$$

where $\lambda_{cont}, \lambda_{sld}, \lambda_{ort}, \lambda_{smt}$ are coefficients of loss terms. $\mathcal{L}$ is minimized with a gradient descent algorithm to iteratively optimize $\boldsymbol{M}^{(i)} = (T^{(i)}, \theta^{(i)})$, where (i) indicates the iteration. $\boldsymbol{M}^{(0)}$ is set as (T^W, θ^I),

Plane detection. To improve validity of foot contact, we detect the planes near the human. We first use Cloth Simulation Filter (CSF) [53] to extract ground points $\boldsymbol{S}_g$ in $\boldsymbol{S}$. And then we search neighboring points of T^W in $\boldsymbol{S}_g$. Unlike the dense mesh model, the discrete point cloud has empty areas, resulting in invalid foot contact constraints. To adress this, we use RANSAC [31] to fit planes for the neighboring points. We denote the plane fuction as p_k.

Foot contact constraint. The foot contact loss is defined as the distance from a stable foot to its nearest ground. Unlike

HPS knowing the information about which foot is stepping on the floor, we detect the foot state based on the movements. First, we compare the left and right foot movement for every successive foot vertices in $V^I_k = \Phi(T^I_k, \theta^I_k, \beta)$ from IMU. One foot is marked as a stable foot if its movement is smaller than $2cm$ and smaller than another foot's movement. The k-th frame's stable foot vertices index list in V_j is denoted as $\mathcal{S}_k$ and the foot contact loss $\mathcal{L}_{cont}$ is written as:

$$\mathcal{L}_{\text{cont}} = \frac{1}{l}\sum_{j=k}^{k+l}\sum_{v_c\in V^{\mathcal{S}_j}_j}\frac{1}{|\mathcal{S}_j|}\|v_c - \widetilde{v}_c p_j\|_2, \quad (3)$$

where $\widetilde{v}_c$ is homogeneous coordinate of v_c. $V^{\mathcal{S}_j}_j$ is denoted as $\mathcal{S}_j$ foot vertices in $V_j = \Phi(\boldsymbol{M}_j)$, which is from the motion to be optimized.

Foot sliding constraint. The foot sliding constraint reduces the motion's sliding on the ground, making the motion more natural and smooth. The sliding loss is defined as every two successive stable foot's distance:

$$\mathcal{L}_{\text{slid}} = \frac{1}{l}\sum_{j=k}^{k+l-1}\|\mathbb{E}(V^{\mathcal{S}_{j+1}}_{j+1}) - \mathbb{E}(V^{\mathcal{S}_{j+1}}_j)\|_2, \quad (4)$$

where $\mathbb{E}(\cdot)$ is the average function.

Orientation constraint. This constraint encourages the motion $\boldsymbol{M}$ to rotate as smooth as IMU and have the same orientation with the landmarks $\mathcal{A}$ described in Sec. 4.4. The orientation loss is written as follows:

$$\mathcal{L}_{\text{orit}} = \frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}\|(R_j)^{-1}R^W_j\|_2 +$$
$$\frac{1}{l-1}\sum_{j=k}^{k+l-1}\max(0, \|(R_j)^{-1}R_{j+1}\|_2 - \|(R^I_j)^{-1}R^I_{j+1}\|_2).$$
$$(5)$$

Smooth constraint. This constraint encourages the human motion to move as smoothly as IMU motion, minimizing the difference between LiDAR and IMU sensors' translation distance. The smooth loss term is as follows:

$$\mathcal{L}_{smt} = \frac{1}{l}\sum_{j=k}^{k+l-1}\max(0, \|T_k - T_{k+1}\|_2 - \|T^I_{k+1} - T^I_k\|_2).$$
$$(6)$$

5. Experiments

This section introduces our dataset and evaluates HSC4D in large indoor-outdoor 3D scenes. The results demonstrate the effectiveness and the generalization ability of HSC4D.
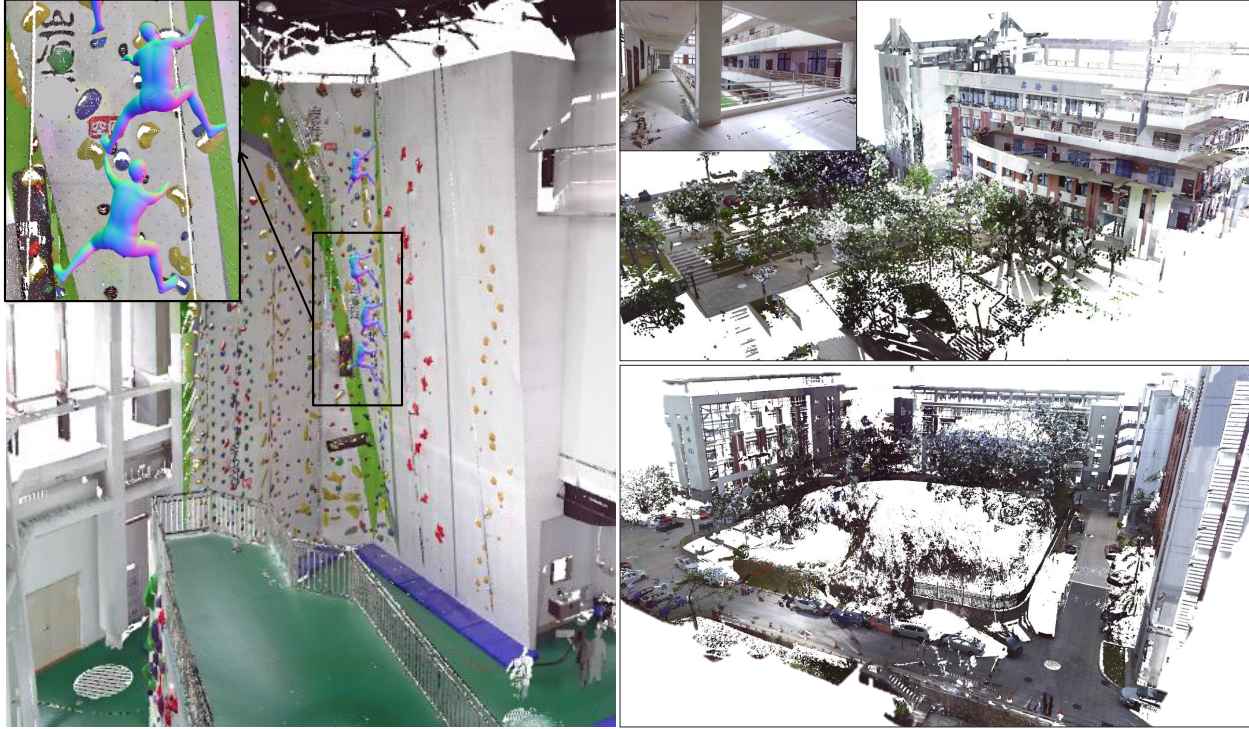
Figure 5. The large indoor and outdoor scenes in our dataset. **Left**: a climbing gym (1200 m^2). **Right top**: a lab building with an outside courtyard 4000 m^2. **Right bottom**: a loop road scene 4600 m^2.

5.1. Dataset

We propose an HSC4D dataset containing three large scenes: a rock climbing gym, a multi-story building, and an outdoor closed-loop road. The gym has a wall height of 20 meters, with a ground and climbing area size over 1200 m^2. The building scene's indoor and outdoor area size is up to 5000 m^2. The scenes have a diversity of heights and environments, including multi-story, slope, and staircase. The outdoor closed-loop road is $70m \times 65m$ with slope. In these scenes, captured human activities include walking, exercising, walking up/down stairs, rock climbing, speeching, etc. Since the 3D map from LiDAR lacks color information, we use a Terrestrial Laser Scanner (Trimble TX5) to scan the color scenes for better visualization. In summary, The HSC4D dataset provides 250K IMU frames (100Hz), 50k time-synchronized LiDAR frames (20Hz), our SLAM results, and colored ground truth point clouds of the scenes.

	HPS	HSC4D
Sensors	IMU-based Mocap Head-mounted camera(*) **LiDARs with cameras(+)**	IMU-based Mocap Hip-mounted LiDAR(*)
Input	IMU motions Video frames(*) **Pre-built 3D map(+)**	IMU motions LiDAR frames(*)

Table 1. **Comparison between HSC4D and HPS.** (*): Different object. (+): Additional object.

5.2. Comparison

The most related work to ours is HPS [9], which uses IMUs, a head-mounted camera, and NavVis M6 equipped with 6 cameras and 4 LiDARs for human-scene modeling. Besides, HPS also heavily records all images registered to the captured 3D map for localization. In contrast, We remove the tedious reliance on the pre-built map in HPS abtained using NavVis. Our approach is scene-prior-free and contributes a novel body-worn setting of LiDAR and IMUs for more practical human-scene modeling. (Tab. 1)

Baselines. Since HSC4D is the first method in such scenes, there is no published baseline available to compare. For effective comparison, we name the IMU result as **Baseline1** and IMU + LiDAR without any optimization as **Baseline2**.

Global localization comparison. First, in the ground truth point clouds provided by Trimble TX5, we mark some locations on the ground as checkpoints. Then the subject steps on the checkpoint during capturing. At last, we measure the distance from the SMPL model estimated by the method to the checkpoint as global localization error.

Local pose comparison. To evaluate the local pose accuracy and smoothness, we compare the foot contact loss $\mathcal{L}_{cont}$ and the sliding loss $\mathcal{L}_{sld}$ described in Sec. 4.5. The comparison is shown in Tab. 4.

Tab. 3 shows the localization error comparisons between our method and baselines. As the distance increases in the

Sequences	$r=0.6$m $l=100$	Loss term			Scene cropping radius			Length of optimization	
		w/o $\mathcal{L}_{cont}$	w/o $\mathcal{L}_{sld}$	w/o $\mathcal{L}_{smt}$	$r=0.4$m	$r=0.8$m	$r=1.0$m	$l=50$	$l=200$
Road	0.96/1.06	+4.44/-0.03	+0.01/+1.06	-0.06/+0.43	+0.02/+0.00	-0.01/+0.00	-0.01/+0.00	-0.03/-0.06	-0.02/-0.03
Gym01	0.66/0.65	+7.03/-0.06	+0.02/+0.70	-0.31/-0.05	+0.02/+0.00	-0.01/+0.00	-0.01/-0.01	+0.00/+0.00	+0.01/+0.00
Building01	0.80/1.20	+5.47/-0.21	+0.04/+0.77	-0.10/-0.35	+0.04/-0.12	+0.02/-0.13	+0.01/-0.13	-0.03/-0.22	+0.02/-0.13
Building04	0.92/0.96	+5.88/-0.04	-0.04/+0.86	-0.12/-0.40	+0.03/+0.00	-0.01/+0.00	-0.02/+0.01	-0.02/+0.01	-0.01/+0.00

Table 2. **Quantitative evaluation of our method:** We record $\mathcal{L}_{cont}$ and $\mathcal{L}_{sld}$ in column two. Column 3-5: loss term ablating comparisons. Column 6-8: analysis of neighborhood radius r for cropping the scene. Column 9-10: analysis of the optimization sequence length l.

CheckPoint	Distance to start	Baseline1	Balseline2	HSC4D
Building (P_1)	$35m$	65.96	8.14	**7.01**
Building(P_2)	$350m$	214.74	17.79	**11.02**
Building(P_3)	$480m$	331.84	24.45	**18.11**
Road(P_{end})	$250m$	192.17	89.15	**67.93**

Table 3. **Global localization error comparison between baselines and HSC4D**. **Baseline1**: IMU result. **Baseline2**: IMU pose + LiDAR localization. Error is measured as distance (cm) from the selected CheckPoint (CP) to the SMPL model's foot.

Terrain	Sequences	Baseline1	Baseline2	HSC4D
		L_{cont}/L_{sld}	L_{cont}/L_{sld}	L_{cont}/L_{sld}
Flat	Building01	2.35/**1.00**	6.22/2.57	**0.80**/1.20
Flat	Building03	3.54/**0.80**	4.13/2.26	**0.64**/0.91
Flat	Gym02	4.20/0.43	5.33/1.08	**0.78/0.20**
Flat	Gym03	44.156/1.14	3.21/2.08	**0.79/0.26**
Flat	Gym01	187.65/**0.56**	7.88/1.72	**0.66**/0.65
Stairs	Building02	389.70/1.05	7.58/2.67	**0.81/0.87**
Stairs	Building04	394.46/**0.89**	6.80/3.09	**0.92**/0.96
Slope	Road	445.39/**1.03**	8.59/3.70	**0.96**/1.06

Table 4. **Local pose error comparison between baselines and HSC4D**. *Building01/03*: 1.5min/1min sequence on the second/first floor's corridor. *Building02/04*: 3min/1min outdoor sequences, including stairs and slope. *Gym01/02/03*: three one-minute walking and warming up sequences. *Road*: 3.5 minute walking sequence.
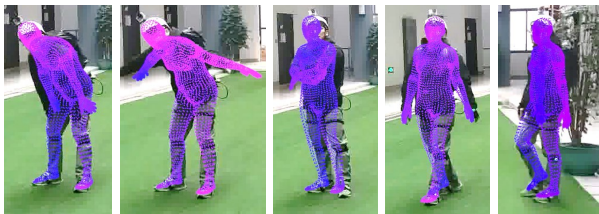

Figure 6. Qualitative results of local pose projected to images.

scene, the error increases linearly in all methods. Baseline1's error is ten times compared to other methods because IMU drifts severely over time. Baseline2 has a smaller global localization error, but its accumulative errors still vary from $8cm$ to $90cm$. The last column shows that HSC4D achieves the smallest global localization errors in the multi-story building and the road with slopes. More specifically, HSC4D improves 78.3% accuracy compared to Baseline1 and 25.4% compared to Baseline2. In Fig. 7, we show comparisons between baselines and our HSC4D.
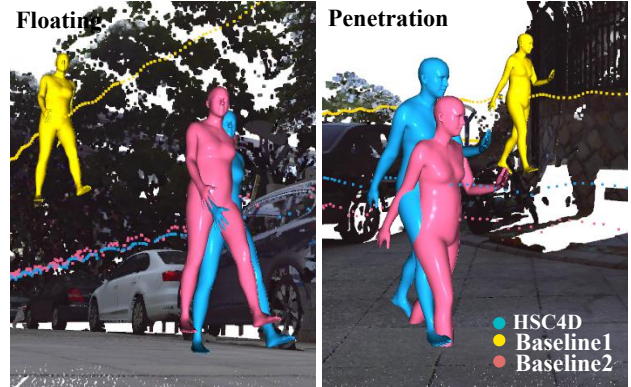

Figure 7. **An example of comparison between Baseline1, Baseline2, and HSC4D.** The colorful spheres represent the trajectories. HSC4D results are natural and accurate in both cases. Left: Baseline1 and Baseline2 are floating in the air. Right: Baseline1 is floating in the air, and Baseline2's lower leg penetrates the ground.

Tab. 4 shows the comparison of local pose errors between our method and baselines. Baseline1's foot contact loss is much larger than other methods especially in scenes where height change. Baseline2's $\mathcal{L}_{sld}$ is the largest among all methods. In the first three sequences where Baseline1 is not drift over height, Baseline2's $\mathcal{L}_{cont}$ is much more larger than Baseline1. These cases indicate that LiDAR increases local errors. See from the last column, HSC4D significantly decreases $\mathcal{L}_{cont}$ in all cases and achieve comparable smoothness on $\mathcal{L}_{sld}$ compared to Baseline1. These comparisons reveal that HSC4D can achieve smooth results in local pose and is robust in various height diversity scenes.

5.3. Evaluation

Quantitative evaluation. We evaluate our optimization method with ablating different loss terms by analyzing $\mathcal{L}_{cont}$ and $\mathcal{L}_{sld}$. We also analyze two parameters of the optimization: the neighborhood radius r, which is used to crop the scene for foot contact constraint, and the subsequence length l of optimization. As shown in Tab. 2, without the foot contact constraint or the foot sliding constraint, $\mathcal{L}_{cont}$ or $\mathcal{L}_{sld}$ increases dramatically and has little impact on another term. Both $\mathcal{L}_{cont}$ and $\mathcal{L}_{sld}$ decrease without the smoothness constraint. However, the motion jitters more severely in 3D visualization. Overall, all loss terms are necessary to produce accurate and smooth human motion. We

Figure 8. Qualitative results of human activities in diverse scenes. The SMPL models in each figure are from the same human subject.

observe that both r and l have little impact on the result. In practice, to balance the computational resource and running time, we set $l = 100$. And to ensure we can detect points in the scenes, we set $r = 0.6m$.

Qualitative evaluation. For local pose qualitative evaluation, we use an extra camera registered to the scene and then project the human motions to the images Fig. 6. We show more qualitative examples in Fig. 8. There are various human activities in different indoor and outdoor scenes shown in the figure. The sequence motions shown in figure represent the same human occurring at a successive time. Above examples show that our method can estimate the natural and challenging human motions with global localization.

6. Discussions

Limitations. First, HSC4D uses LiDAR SLAM results as the initial localization. Consequently, it is limited when LiDAR mapping fails, such as crowded places, narrow areas, etc. Second, to avoid a large occlusion during LiDAR data capturing, some activities are limited in our system, such as sitting in a chair with a backrest, standing back against a wall. Besides, the optimization loss terms in our pipeline are hand-created. Some cases like rock climbing do not always work. It is promising to propose a more general loss term or a deep learning framework cooperating with the semantic information.

Conclusions. We present a Human-centered 4D Scene capture method to accurately and efficiently create a dynamic digital world using only body-mounted IMUs and LiDAR. Our method is space-free, pose-free, and map-free. By integrating LiDARs and IMUs, our proposed joint optimization algorithm can obtain accurate global localization and smooth local poses in large scenes. Additionally, we provide a new dataset containing large scenes and diverse, challenging human motions. The experimental result demonstrates the effectiveness of HSC4D. Our work contributes to extending the motion capture to large dynamic scenes. We hope this work will foster the creation and interaction of the human-dynamic digital world in the future.

References

[1] Moustafa Abbas, Moustafa Elhamshary, Hamada Rizk, Marwan Torki, and Moustafa Youssef. Wideep: Wifi-based accurate and robust indoor localization system using deep learning. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*, pages 1–10. IEEE, 2019. 2

[2] Abdulrahman Alarifi, AbdulMalik Al-Salman, Mansour Alsaleh, Ahmad Alnafessah, Suheer Al-Hadhrami, Mai A Al-Ammar, and Hend S Al-Khalifa. Ultra wideband indoor positioning technologies: Analysis and recent advances. *Sensors*, 16(5):707, 2016. 2

[3] 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

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[5] 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

[6] 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. 2

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[8] 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

[9] 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, 6

[10] Thomas Helten, Andreas Baak, Gaurav Bharaj, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. Personalization and evaluation of a real-time depth-based full body tracker. In *International Conf. on 3D Vision*, pages 279–286, 2013. 2

[11] 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. 2

[12] Jianhao Jiao, Haoyang Ye, Yilong Zhu, and Ming Liu. Robust odometry and mapping for multi-lidar systems with online extrinsic calibration. *IEEE Transactions on Robotics*, 2021. 3

[13] 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

[14] 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

[15] 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. 2

[16] Wonhui Kim, Manikandasriram Srinivasan Ramanagopal, Charles Barto, Ming-Yuan Yu, Karl Rosaen, Nick Goumas, Ram Vasudevan, and Matthew Johnson-Roberson. Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters*, 4(2):1940–1947, 2019. 2

[17] Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. g 2 o: A general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation*, pages 3607–3613. IEEE, 2011. 5

[18] Filip Lemic, Jasper Büsch, Mikolaj Chwalisz, Vlado Handziski, and Adam Wolisz. Infrastructure for benchmarking rf-based indoor localization under controlled interference. In *2014 Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS)*, pages 26–35. IEEE, 2014. 2

[19] Qing Li, Shaoyang Chen, Cheng Wang, Xin Li, Chenglu Wen, Ming Cheng, and Jonathan Li. Lo-net: Deep real-time lidar odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8473–8482, 2019. 2

[20] 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

[21] 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

[22] 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, 3

[23] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 2

[24] 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. 2

[25] Helen Oleynikova, Michael Burri, Simon Lynen, and Roland Siegwart. Real-time visual-inertial localization for aerial and ground robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3079–3085. IEEE, 2015. 2

[26] 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

[27] Chao-Chung Peng, Yun-Ting Wang, and Chieh-Li Chen. Lidar based scan matching for indoor localization. In *2017 IEEE/SICE International Symposium on System Integration (SII)*, pages 139–144. IEEE, 2017. 2

[28] 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. 2

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

[30] Cristina Romero-González, Álvaro Villena, Daniel González-Medina, Jesus Martínez-Gómez, Luis Rodríguez-Ruiz, and Ismael García-Varea. Inlida: A 3d lidar dataset for people detection and tracking in indoor environments. In *International Conference on Computer Vision Theory and Applications*, volume 7, pages 484–491. SCITEPRESS, 2017. 2

[31] Ruwen Schnabel, Roland Wahl, and Reinhard Klein. Efficient ransac for point-cloud shape detection. In *Computer graphics forum*, volume 26, pages 214–226. Wiley Online Library, 2007. 5

[32] 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

[33] 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

[34] 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. 2, 3

[35] 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

[36] 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

[37] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4470–4479, 2018. 2

[38] 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. 2

[39] 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

[40] 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. 2

[41] 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. 2

[42] 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. 2

[43] H. Wang, C. Wang, C. Chen, and L. Xie. F-loam : Fast lidar odometry and mapping. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 3

[44] Yun-Ting Wang, Chao-Chung Peng, Ankit A Ravankar, and Abhijeet Ravankar. A single lidar-based feature fusion indoor localization algorithm. *Sensors*, 18(4):1294, 2018. 2

[45] 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

[46] 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

[47] Lan Xu, Wei Cheng, Kaiwen Guo, Lei Han, Yebin Liu, and Lu Fang. Flyfusion: Realtime dynamic scene reconstruction using a flying depth camera. *IEEE transactions on visualization and computer graphics*, 27(1):68–82, 2019. 2

[48] 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

[49] Huan Yin, Li Tang, Xiaqing Ding, Yue Wang, and Rong Xiong. Locnet: Global localization in 3d point clouds for mobile vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 728–733. IEEE, 2018. 2

[50] Shangshu Yu, Cheng Wang, Zenglei Yu, Xin Li, Ming Cheng, and Yu Zang. Deep regression for lidar-based localization in dense urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 172:240–252, 2021. 2

[51] Ji Zhang and Sanjiv Singh. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and Systems*, volume 2, 2014. 2, 3, 4

[52] 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

[53] Wuming Zhang, Jianbo Qi, Peng Wan, Hongtao Wang, Donghui Xie, Xiaoyan Wang, and Guangjian Yan. An easy-to-use airborne lidar data filtering method based on cloth simulation. *Remote Sensing*, 8(6):501, 2016. 5

[54] 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

[55] 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