LiDAR-aid Inertial Poser: Large-scale Human Motion Capture by Sparse Inertial and LiDAR Sensors

Yiming Ren†, Chengfeng Zhao†, Yannan He, Peishan Cong, Han Liang, Jingyi Yu, Lan Xu⋆, and Yuexin Ma⋆

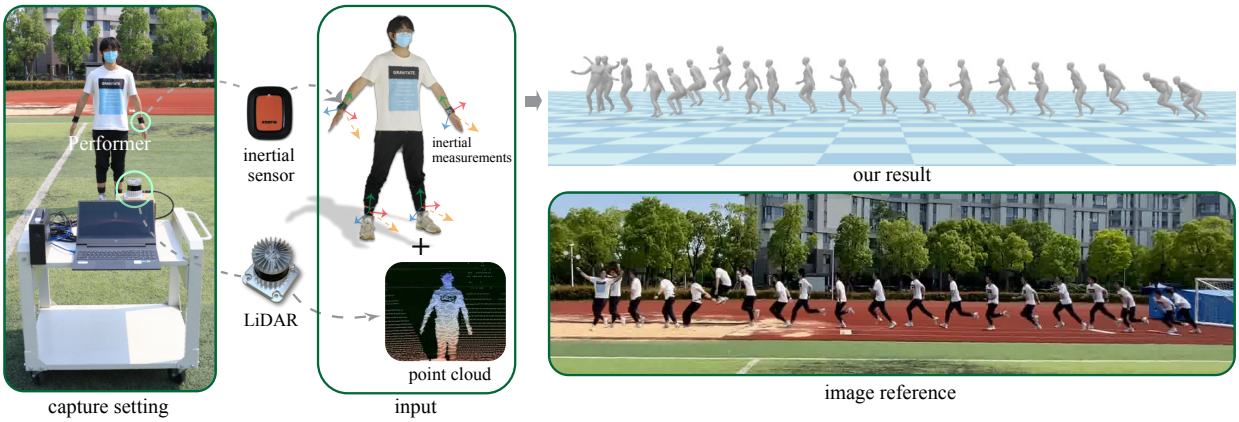


Fig. 1: We propose a multi-modal motion capture approach, LIP, that estimates challenging human motions with accurate local pose and global translation in large-scale scenarios using single LiDAR and four inertial measurement units.

Abstract— We propose a multi-sensor fusion method for capturing challenging 3D human motions with accurate consecutive local poses and global trajectories in large-scale scenarios, only using single LiDAR and 4 IMUs, which are set up conveniently and worn lightly. Specifically, to fully utilize the global geometry information captured by LiDAR and local dynamic motions captured by IMUs, we design a two-stage pose estimator in a coarse-to-fine manner, where point clouds provide the coarse body shape and IMU measurements optimize the local actions. Furthermore, considering the translation deviation caused by the view-dependent partial point cloud, we propose a pose-guided translation corrector. It predicts the offset between captured points and the real root locations, which makes the consecutive movements and trajectories more precise and natural. Moreover, we collect a LiDAR-IMU multi-modal mocap dataset, LIPD, with diverse human actions in long-range scenarios. Extensive quantitative and qualitative experiments on LIPD and other open datasets all demonstrate the capability of our approach for compelling motion capture in large-scale scenarios, which outperforms other methods by an obvious margin. We will release our code and captured dataset to stimulate future research.

Index Terms—Human motion capture, shape modeling, virtual reality, sensor fusion

1 Introduction

With the rise of VR/AR over the last decades, human motion capture (mocap) evolves as a cutting-edge technique for humans to interact and communicate with each other in virtual worlds using our body language. Despite the huge progress of data-driven mocap solutions, the accurate and convenient capture of human motions in large-scale scenarios remains challenging. It is critical for the reconstruction, simulation, and generation of sporting mega-events, stage performances, interactions of crowds, etc., and has recently received substantive attention.

By far, optical-based solutions take the majority of human mocap. The high-end marker-based solutions [44, 61, 62] require outside-in multi-camera setup or dense optical markers, and thus confine the performers to a constrained captured area, making large-scale capture impractical. Recently, learning-based monocular methods [14, 15, 24, 25, 27, 28, 30, 50, 77] enable robust motion capture under light-weight setting. Although alleviating expensive facilities and fixed captured region, they remain vulnerable to occlusions, lack of textures and severe changes of environment lighting, etc,. Moreover, their inherent lack of depth measurement makes them unstable to accurately track global trajectories of humans, especially under large-scale settings.

In contrast, motion capture using inertia information recorded by Inertial Measurement Units (IMUs) is occlusion-unaware and environment-independent. The high-end solutions [43, 68] require a large amount of body-worn IMUs (from 8 to 17), making them unsuitable to capture human motions with everyday apparel. Recent data-driven advances [19, 73, 74] enable real-time motion capture with sparse IMUs. They can obtain global location by accumulating the IMU observation and even partially alleviate the drifting with foot-ground contact or physical constraints. Yet, such purely inertial methods still inherently suffer from temporally cumulative error of global localization, especially for capturing large-scale motions with a long duration. Only recently, LiDAR-based mocap is gaining increasing attention due to the tremendous progress of LiDAR in large-scale perception [7, 8, 81, 82]. Notably, LiDARCap [32] leverages a graph-based convolutional network to predict daily human poses from the point clouds captured by a LiDAR sensor within range of 30 meters. However, it's fragile to capture challenging human motions or multi-person scenes with strong

- †: *Equal contribution.*
- ⋆: *Corresponding author.*
- *Yiming Ren, Chengfeng Zhao, Yannan He, Peishan Cong, Han Liang are with ShanghaiTech University. E-mail: { renym2022 | zhaochf2022 | heyn | congpsh | lianghan | yujingyi | xulan1 | mayuexin}@shanghaitech.edu.cn.*
- *Jingyi Yu, Lan Xu, and Yuexin Ma are with ShanghaiTech University and Shanghai Engineering Research Center of Intelligent Vision and Imaging. E-mail: {yujingyi | xulan1 | mayuexin}@shanghaitech.edu.cn.*

Manuscript received 14 October 2022; revised 13 January 2023; accepted 30 January 2023. Date of publication 22 February 2023; date of current version 29 March 2023. Digital Object Identifier no. 10.1109/TVCG.2023.3247088

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.
1077-2626 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

self-occlusions and external occlusions due to the view-dependent and sparsity-varying properties of LiDAR point clouds.

To tackle the above challenges, we propose a novel approach called LiDAR-aid Inertial Poser (LIP), to capture challenging human motions in large-scale scenarios accurately. In stark contrast with previous mocap system, our LIP adopts a novel and light-weight hardware setup using only single LiDAR and 4 IMUs. These two kinds of sensors are complementary to each other, providing both global geometry and local dynamics information of the captured performer. Meanwhile, they are naturally privacy-preserving and insensitive to the lighting, which is appropriate to be generalized to novel scenes. As shown in Fig. 1, our system faithfully reconstructs both local skeletal poses and global trajectories of the performer under a novel sensor-fusion setting.

Generating robust mocap results using such novel multi-modal inputs is non-trivial. First, point cloud only capture the global visual information, while IMU measurements encode the local actions physically. Furthermore, the partially captured point clouds of human body will significantly change under various poses and views of the performer, which negatively affects the accurate localization and naturally consecutive motion capture. To this end, in LIP, we adopt two consistent technical modules to accurately estimate human poses and translations, respectively. For the former one, we introduce a **Multi-modal Pose Estimation** module, which is two-stage, conducting sensor fusion and local body motion estimation in a coarse-to-fine manner. Specifically, we first extract pose and motion patterns from point cloud by regressing 24 joints of the parametric human model SMPL [35] and the 6D rotation of the root joint through a convolutional temporal encoder (PointNet-GRU) in Stage1. Next, a hierarchical pose estimation process with fused IMU measurements is appended to refine the coarse skeleton by a joint-map estimator and solve the inverse kinematics(IK) problem for the joint rotations by a body-pose estimator in Stage2. To precisely capture the 3D global motion trajectories, we introduce the second module called **Pose-guided Translation Correction**, so as to learn the deviation between the partially captured point could and the real location of the root joint. Specifically, considering the fact that diverse poses can affect the deviation, we use another isomorphic PointNet-GRU structure to model the shape and pose characteristics from point cloud, cooperated with the estimated joint rotations and refined skeleton. We further utilize the fused pose feature to eliminate the inherent translation deviation and random noise from the captured point cloud.

In particular, to facilitate the research of multi-modal human mocap tasks, we collect a huge LiDAR-IMU hybrid mocap dataset, LIPD, consisting of rich and challenging single-person and multi-person actions in long-range scenes. It is the first dataset to provide multi-modal LiDAR point cloud and IMU measurements and ground-truth SMPL pose parameters for the mocap task. We conduct extensive experiments on LIPD and a variety of other real and synthetic datasets [19,32,33,38] to demonstrate the capability of our method LIP for capturing challenging human motions in various large-scale scenes. To summarize, our main contributions include:

- We propose the first LiDAR-IMU hybrid approach for 3D human motion capture in large-scale scenarios and achieves state-of-the-art performance.
- We propose an effective two-stage coarse-to-fine fusion method to fully utilize the complementary features of multi-modal input for accurate pose estimation.
- We propose an approach to eliminate the translation deviation in a pose-guided manner, achieving accurate global trajectories and natural consecutive actions.
- We provide a huge LiDAR-IMU-based multi-modal mocap dataset with diverse human actions in various large-scale scenarios, which can benefit both single modal and multi-modal mocap research works.

2 Related Work

Optical Motion Capture. Marker-based motion capture studios [44, 61, 62] enable capturing high-quality professional motions, which have achieved success and are widely used in the industry. However, these systems are costly and cumbersome, and performers usually need to wear the marker suits, which means unavailable for daily usage. To overcome these problems, the exploration of markerless mocap [5,9,17,22,37,55,56,58,59,69,70,76] has made great progress. Previous multi-view algorithms [1, 6, 10, 47, 51, 52, 57] demonstrate robust motion capture even in the wild. Although the cost and intrusiveness is drastically reduced, synchronizing and calibrating multi-camera systems is still tedious. Thus various monocular mocap approaches have been proposed, which estimate 3D human pose and shape by optimizing [4, 18, 29, 31] or directly regressing [24, 25, 27, 77]. To overcome various flaws of monocular setup, template-based approaches [13–15,71,72], probabilistic approaches [30,50] and semantic-modeling approaches [28] are proposed. However, the inherent depth ambiguity is still unsolved. Some approaches [3, 12, 54, 66, 75] using the commodity depth cameras enable alleviating this, but these active IR-based cameras are unsuitable for outdoor capture and the capture volume is limited. Recently, Li et al. [32] employs a consumer-level LiDAR which provides large-scale depth information, to enable large-scale 3D human motion capture. However, LiDAR point cloud is view-dependent and sparsity-varying and the pure LiDAR-based method suffers from severe occlusions, which hinders the generalization to more challenging human actions and multi-person scenarios where occlusions are inevitable.

Inertial Motion Capture. In contrast to optical approaches based on cameras, purely inertial approaches using IMUs are free from occlusion and restricted recording environment and volume. However, commercial purely inertial solutions such as Xsens MVN [68] rely on large amounts of IMUs. Performers are usually required to wear tight suits with densely bounded IMUs, which is intrusive, tedious and inconvenient, and prompt the community forward to the sparse setup. A pioneering work, SIP [65], presents an optimization-based offline method using only 6 IMUs and achieves promising results. Inspired by it, recent data-driven approaches [19, 73, 74] with the same setup achieve great improvements in accuracy and efficiency, which enable real-time pose and translation estimation. Nevertheless, these purely inertial approaches still suffer from substantial drifts while performing high-speed or large-scale capture.

Hybrid Motion Capture. As optical and inertial mocap solutions suffer from occlusions and drifts, respectively. Approaches that fuse these two types of sensors to benefit from complementarity, so as to achieve more robust mocap, have attracted much attention. Preceding methods propose to combine IMUs with RGB cameras, which can be achieved by either optimization [23, 39–41, 48] or regression [11, 60, 64, 78]. Recently, Liang et al. [34] presents a learning-and-optimization method fusing a single camera with only 4 IMUs, which demonstrates robust challenging motion capture. Besides, some works fuse IMUs with depth cameras [16, 79] or optical markers [2], which achieve promising results. To improve interactive and immersive experiences in AR/VR applications, motion tracking methods using head mounted devices(HMD) mixed with other sparse sensors bring great advances [21, 67]. Nevertheless, these methods still suffer from limited capture volume and are sensitive to light, which limits the usage for large-scale motion capture. Furthermore, following traditional optimization schemes, several methods [46, 83] combine dense IMUs with multiple LiDARs, which is inconvenient to set up and too heavy for performers. In this work, we propose a novel data-driven hierarchical mocap approach by fusing only single LiDAR and 4 IMUs, which is lightweight and alleviates both occlusion and drift problems and achieves accurate 3D challenging motion capture in large-scale scenes.

Motion Capture Dataset. Current booming data-driven mocap methods rely on huge labeled datasets. There are already many open datasets for facilitating related research. Human3.6M [20] is a popular and wildly used motion capture dataset, which records 3.6 million frames data by 10 high speed motion cameras, it covers 17 daily motions performed by 11 actors, and a similar dataset HumanEva [55]

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.

provides 6 common actions. Both of them are marker-based datasets and rely on surrounding multi-view cameras to gather 3D human poses, thus the activity space, clothing, actions and lighting conditions are limited. AMASS [38] integrates them into a single meta dataset with a common framework and parameterization for the training of data-driven networks. MPI-INF-3DHP [42] captures ground-truth from commercial marker-less motion capture in a multi-camera green screen studio and AIST++ [33] captures ground-truth by 3D reconstruction method from multi-view videos. Such marker-less datasets cover more complex poses than marker-based ones. TotalCapture [60] is the first dataset to utilize IMU to capture the pose, which consists of videos and corresponding 13 IMU sensor measurements. All above datasets are limited to indoor scenes. As the demand of mocap in outdoor large-scale scenes grows, some outdoor datasets emerged. PedX [26] is a large-scale outdoor 3D datasets captured by stereo cameras and LiDAR, but the full 3D labels are generated by 2D annotations. 3DPW [63] gathers motion ground-truth by hand-held smartphone camera and IMU sensors, and DIP-IMU [19] uses 17 IMU sensors to generate ground-truth by SIP. Nevertheless, both camera and IMUs can not provide accurate depth information, making these methods difficult to reconstruct the human motions in the whole scene. LiDARHuman26M [32] proposes the first large-scale benchmark dataset featuring LiDAR point cloud and IMU-captured human motion ground truth. It utilizes LiDAR point cloud as pure input to acquire the global localization information and eliminate the local pose ambiguity. However, the dataset does not provide raw IMU measurements and only consists of 20 simple daily single-person motions with few self-occlusion cases. We propose a LiDAR and IMU multi-sensor-based mocap dataset with diverse challenge human actions in various large-scale scenarios, which is applicable for both general and professional scenarios.

3 Method

Our goal is to capture challenging 3D human motions in large-scale scenarios with consistently accurate local pose and global trajectory estimation using single LiDAR and 4 IMUs. Fig. 2 illustrates an overview of the entire pipeline, which consists of two cooperative modules to infer pose and translation, respectively. For pose inference module, we propose a two-stage network working in a coarse-to-fine manner to distill rough human body skeleton and global orientation from raw point cloud first, and then regress the human body pose from joint positions refined by IMU measurements (Sect. 3.2). For translation inference module, we design a pose-guided approach to learn the latent domain gap between LiDAR measurements and real global movements, through which the inherent translation deviation and random noise from point cloud can be eliminated (Sect. 3.3).

3.1 Preliminaries

In this section, we give detailed explanations for the design of our network. Before that, we clarify our system input pre-processing and the most frequently used math symbols in the following.

System Input Pre-processing: Since raw point cloud sequence with variable $N_p(t)$ points at different time frame t is temporally inconsistent, we normalize the point cloud in every single frame $\widetilde{\mathbf{x}}^{(1)}(t) \in \mathbb{R}^{N_p(t)\times 3}$ to $\mathbf{x}^{(1)}(t) \in \mathbb{R}^{N_{fps}\times 3}$ by subtracting its arithmetic mean and resampling to fixed N_{fps} points using farthest point sampling algorithm (FPS). In our implementation, we set $N_{fps} = 256$. For IMU measurements, we transform sensors' inertia in inertial coordinate system to bones' inertia in LiDAR coordinate system, and formulate single-frame inertial input as $\mathbf{x}^{(2)}(t) = [\mathbf{R}_{lw}, \mathbf{R}_{rw}, \mathbf{R}_{la}, \mathbf{R}_{ra}, \mathbf{a}_{lw}, \mathbf{a}_{rw}\mathbf{a}_{la}, \mathbf{a}_{ra}] \in \mathbb{R}^{48}$, where $\mathbf{R}$ denotes the rotation in flattened rotation matrix form while $\mathbf{a}$ indicates the free acceleration, and lw, rw, la, ra mean left wrist, right wrist, left ankle, and right ankle, respectively.

Definition of Motion Representations: We define N_j, N_s as the amount of body joints and IMU sensors; $\hat{\mathbf{J}}(t)$, $\mathbf{J}^{GT}(t)$, $\widetilde{\mathbf{J}}^{GT}(t) \in \mathbb{R}^{3N_j}$ as predicted root-relative joint positions, ground-truth root-relative joint positions, and ground-truth absolute joint positions at time t; $\hat{\mathbf{\Theta}}(t)$, $\mathbf{\Theta}^{GT}(t) \in \mathbb{R}^{6N_j}$ as predicted joint rotations and ground-truth joint rotations in 6D rotation representation [80] at time frame t. Note that all the formulations of loss functions defined below omit a common factor $\frac{1}{T}$ where T is the total time length of training motion sequences.

3.2 Multi-modal Pose Estimation

Purely inertial or LiDAR-based methods more or less suffer from insufficient observations. First, without global visual cues, the reconstruction from local inertia to the 3D joint position is ambiguous and the global location is also inaccurate. Secondly, the view-dependent LiDAR point cloud lacks the representation of unseen body parts, causing difficulties in motion prediction when severe occlusions occur. Therefore, we propose to estimate body pose with multi-modal input, which includes both global geometry information from the LiDAR point cloud and local dynamic motion information from inertia measurements. However, since point cloud is in spatial form while IMU measurement is a physical quantity, the huge domain gap makes it irrational to directly concatenate $\mathbf{x}^{(1)}(t)$ and $\mathbf{x}^{(2)}(t)$ as network input directly. Instead, we formulate this module as a two-stage network working in a coarse-to-fine manner so that the point cloud can combine with motion inertia efficiently. *Global Temporal Pose-prior Distillation* is designed to extract hidden geometric feature and motion pattern from the normalized point cloud and express it explicitly. After that, *Hierarchical Pose Estimation* fuses IMU measurements in and regresses body joint rotations from refined 3D body joint positions.

Global Temporal Pose-prior Distillation: Considering that the raw point cloud can directly represent the coarse human shape and pose information and sparse IMUs can not directly estimate the root joint orientation, we distill the human-skeleton-joint positions and the root orientation from the point cloud in the first stage and then use the four IMU sensors in the arms and legs to refine the result and regress the shape parameters. Specifically, We propose a global temporal pose-prior distillation to regress the 24 SMPL joint positions and the 6D root orientation, which is composed of a global feature extractor PointNet [49] and a temporal encoder two-way GRU(bi-GRU). PointNet is used as encoder to extract human skeleton geometric information from the raw point cloud as a feature vector $v(t) \in \mathbb{R}^k$ from each frame $F(t)$, where k = 1024. We feed the frame-wise features $v(t)$ into the two-way GRU(bi-GRU) to generate the hidden variables $h(t)$ to extract temporal information. Then, we use the MLP decoder to predict the joint positions $\hat{\mathbf{J}}_{prior}(t)$ and the root orientation $\hat{\mathbf{\Theta}}_{root}(t)$, which forms part of the second stage of input. We extract 24 SMPL body joints $\mathbf{J}^{GT}(t)$ and select 6D root rotation from SMPL pose parameters $\mathbf{\Theta}^{GT}_{root}(t)$ as the ground truth. The losses of the above two supervision information can be formulated as

$$\mathcal{L}_{joint-prior} = \sum_t \| \hat{\mathbf{J}}_{prior}(t) - \mathbf{J}^{GT}(t) \|_2^2, \tag{1}$$

$$\mathcal{L}_{ori-prior} = \sum_t \| \hat{\mathbf{\Theta}}_{root}(t) - \mathbf{\Theta}^{GT}_{root}(t) \|_2^2, \tag{2}$$

$$\mathcal{L}_{prior} = \lambda_1 \mathcal{L}_{joint-prior} + \lambda_2 \mathcal{L}_{ori-prior}, \tag{3}$$

where λ_1 and λ_2 are hyper-parameters.

Hierarchical Pose Estimation: Although the explicit pose-prior distilled from the point cloud is available for a neural Inverse Kinematics (IK) solver, it is too coarse to give accurate 3D joint positions and robust inference for challenging motions due to the lack of locally precise motion observations. Therefore, we hierarchically organize this stage as two estimators, which refine the root-relative 3D joint positions with the aid of IMU measurements and then regress the joint rotations. These two estimators share the same architecture, composed of three GRUs connected sequentially with a skip connection, which can make use of temporal motion information and maintain a healthy back propagation for our deep network during training. The former one, *Joint-map Estimator*, takes $\mathbf{x}^{(3)}(t) = [\mathbf{x}^{(2)}(t), \hat{\mathbf{J}}_{prior}(t), \hat{\mathbf{\Theta}}_{root}(t)] \in \mathbb{R}^{12N_s+3N_j+6}$ as the input and outputs a refined root-relative 3D joint-map with more accurate positions $\hat{\mathbf{J}}_{fine}(t)$, supervised by the loss function

$$\mathcal{L}_{fineJ} = \sum_t \| \hat{\mathbf{J}}_{fine}(t) - \mathbf{J}^{GT}(t) \|_2^2 . \tag{4}$$

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.

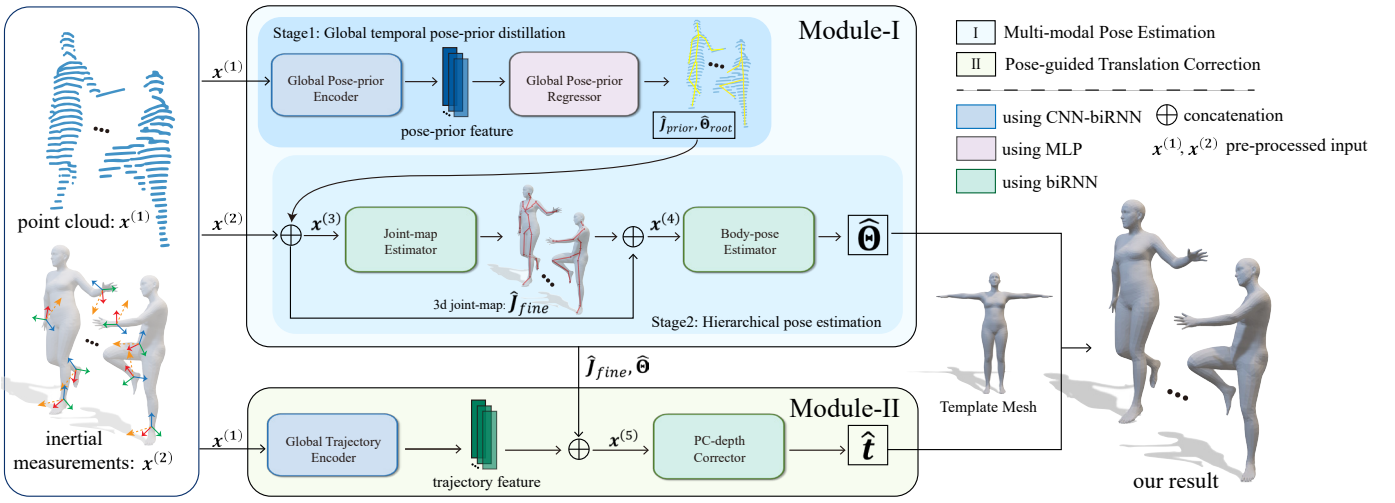


Fig. 2: Overview of our pipeline. It consists of two cooperative modules: *Multi-modal Pose Estimation*(Module-I) and *Pose-guided Translation Correction*(Module-II) to estimate skeletal pose and global translation from sparse IMU and LiDAR inputs.

After that, $\mathbf{x}^{(4)}(t) = [\mathbf{x}^{(3)}(t), \hat{\mathbf{J}}_{fine}(t)] \in \mathbb{R}^{12N_i+3N_j+3N_j+6}$ is fed into *Body-pose Estimator* which learns to solve an IK-like problem to regress joint rotations $\hat{\mathbf{\Theta}}(t)$, supervised by the loss function

$$\mathcal{L}_{ik} = \sum_t \| \hat{\mathbf{\Theta}}(t) - \mathbf{\Theta}^{GT}(t) \|_2^2 . \tag{5}$$

Because of the limited expression capability of joint rotations, we introduce the loss Equation 6 for *Body-pose Estimator* by Forward Kinematics (FK) to better reconstruct the 3D joint positions.

$$\mathcal{L}_{fk} = \sum_t \| \mathrm{FK}(\hat{\mathbf{\Theta}}(t)) - \mathbf{J}^{GT}(t) \|_2^2 . \tag{6}$$

The complete loss function of this module is formulated as

$$\mathcal{L}_{pose} = \lambda_3 \mathcal{L}_{fineJ} + \lambda_4 \mathcal{L}_{ik} + \lambda_5 \mathcal{L}_{fk}. \tag{7}$$

3.3 Pose-guided Translation Correction

Global translation estimation, especially in the large-scale scenario, is challenging for purely inertial mocap methods since no localization information can be provided by IMUs. Moreover, IMUs suffer from the drifting problem. Even in the state-of-the-art work (TransPose [74]), the capture of global movements should be performed under strong assumptions such as level ground and sufficient foot-ground contacts. LiDAR, however, brings significant benefits to this task by measuring distances directly. LiDARCap [32] utilizes the average of points as global translation, however, there exists a deviation between the scanned point cloud and real movement positions, because LiDAR only collects partial points of the performer in the capture view. Accurately, the deviation differs for different poses. For example, standing still and bending result in similar global translation positions, while different averages of points. Considering this, we construct the relationships between poses and translation deviations and treat the translation discrepancy as a per-pose variable, which is only correlated with the pose performed. In practice, we combine *encoder* and *estimator* block to model the motion pattern from the point cloud and learn this deviation, rather than regress the global translation directly. The ground-truth translation discrepancy $\mathbf{D}^{GT}(t) \in \mathbb{R}^3$ can be calculated as follows:

$$\mathbf{D}^{GT}(t) = \widetilde{\mathbf{J}}^{GT}_{root}(t) - \mathrm{avg}(\widetilde{\mathbf{x}}^{(1)}(t)), \tag{8}$$

where $\widetilde{\mathbf{J}}^{GT}_{root}(t)$ is the ground-truth absolute position of the root joint and the operator avg$(\cdot)$ calculates the approximate center of given point cloud by averaging the positions of all the points. Finally, we use the loss Equation 9 to train this module:

$$\mathcal{L}_{trans} = \sum_t \| \hat{\mathbf{D}}(t) - \mathbf{D}^{GT}(t) \|_2^2, \tag{9}$$

where $\hat{\mathbf{D}}(t)$ denotes the predicted translation discrepancy result.

3.4 Implementation Details

We implement our network using PyTorch 1.8.1 with CUDA 11.1. The training of the two modules is separate, which uses batch size of 32, sequence length of 32, learning rate of 10^{-4} and decay rate of 10^{-4} with AdamW optimizer [36]. The weights of loss functions are: $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.2$, $\lambda_4 = 0.7$, $\lambda_5 = 0.7$. All the training and evaluation are conducted on a server with an Intel(R) Xeon(R) E5-2678 CPU and an NVIDIA RTX3090 graphics card. Furthermore, we use free acceleration input, which escapes from the influence of gravity in inertial coordinate frame.

4 Dataset: LIPD

Table 1: Overview of datasets we use. The data mode of point cloud and IMU measurements for each dataset is reported. Moreover, "Capture distance" shows the maximum distance between performer and capture device; "Activity range" stands for the average 3D space size in which the performer moves. Both of them reflect the scale attribute of each dataset.

Dataset	Point cloud	IMU recording	Capture distance(m)	Activity range(m^3)
DIP-IMU [19]	Syn	Real	N/A	N/A
AMASS [38]	Syn	Syn	3.42	0.27
AIST++ [33]	Syn	Syn	4.23	0.67
LiDARHuman26M [32]	Real	Syn	28.05	142.95
LIPD	Real	Real	30.04	366.34

To learn the local pose and global translation estimator working for large-scale scenarios, a huge LiDAR-IMU hybrid mocap dataset is required, which needs the LiDAR to capture the point cloud data and dense IMU sensors to provide the ground-truth. In this paper, we propose the first long-range LiDAR-IMU hybrid mocap dataset focusing on diverse challenge motions, such as many athletic motions. The dataset contains 15 performers, about 30 kinds of motions and 62,341 frames of LiDAR point cloud and corresponding IMU measurements in total. LIPD also provides ground-truth SMPL pose parameters and RGB images. We divide the data by 39,593 frames as the training set and 22,748 frames as the testing set.

We collect LIPD by Noitom IMU sensors [43] and OUSTER-1 LiDAR [45], performers wear a full set of Notiom equipment to collect themed challenge actions within the range of 12-30m. Furthermore, to show the generalization capability of our method, we record extra 8,193

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.

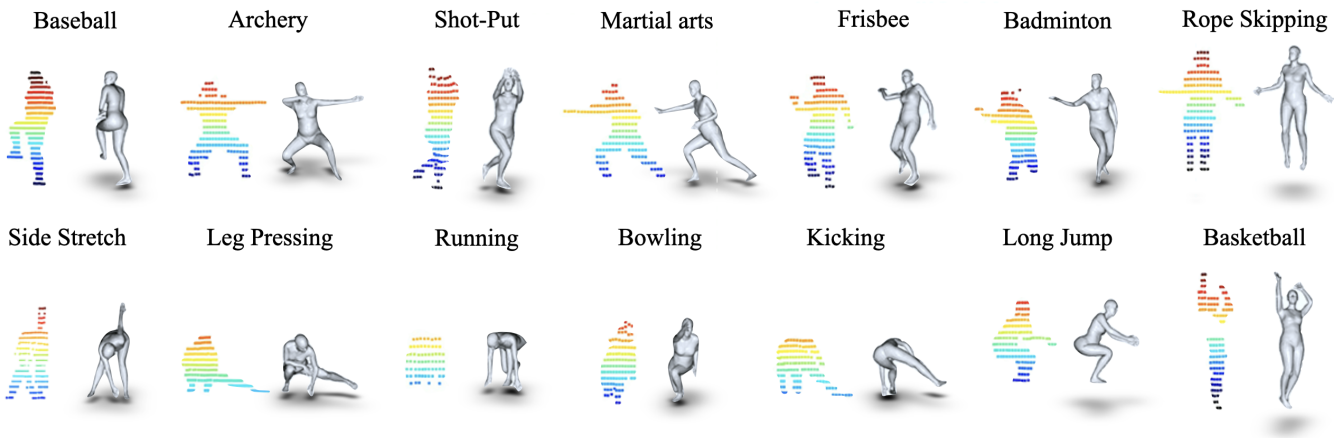


Fig. 3: LIPD dataset contains diverse human poses from daily actions to professional sports actions. We demonstrate the point cloud and corresponding ground-truth mesh for several action samples. Such complex poses, like crouching or having the back to LiDAR in the second row, bring challenges for accurate mocap due to severe self-occlusions. We also have multi-person scenarios with external occlusions, as the last basketball case shows, where the player's body is partially covered by others in the LiDAR's view.

frames of data by our proposed light-weight setting, including single LiDAR and four Xsens Dot IMU sensors [68], in diverse wild scenes for visualization evaluation. Xsens provides professional applications that can record offline IMU data on mobile phones, which means that actors can move freely without space restrictions. Performers only need to wear four Xsens Dot IMU sensors in four limbs for testing, so that they can perform more challenging motions.

4.1 Characteristic.

Large-scale scenes. As mentioned in Table 1, LIPD has $366.34m^3$ activity range, not only providing the long-range data, but also covering scenes in varying heights, such as the climbing motion on stairs. Meanwhile, LIPD provides more views for motion capture compared with LiDARHuman26M [32], which only includes the top view. Furthermore, we also provide the scene in dark to show that the LiDAR can capture accurate point cloud data in extreme environments, which is superior to cameras.

Diverse motions. The LiDAR-only mocap dataset, LiDARHuman26M [32], contains about 20 daily human motions without too many occlusions. Due to the view-dependent property of LiDAR, the performance drops dramatically when handling complex motions with obvious occlusions. LIPD is a new benchmark for mocap in large-scale scenes involving diverse challenging poses, as Fig. 3 shows. It includes both raw LiDAR point cloud and IMU measurements for facilitating the research of accurate mocap in large-scale scenes based on multimodal inputs.

4.2 Extension.

In order to enrich the dataset with more different motions, we simulate plenty of synthetic LiDAR-IMU measurements with ground-truth SMPL [35] pose parameters on DIP-IMU [19], LiDARHuman26M [32], AIST++ [33] and a subset of AMASS [38], including ACCAD, BML-Movi, CMU, and TotalCapture [60]. The synthesized data consists of 74,6267 frames with 4,238 motion genres. As for the protocol of dataset splitting, we take DIP-IMU [19], TotalCapture [60], the testing set of LiDARHuman26M [32], and the testing set of LIPD for evaluation. The left data is for training. Table 1 gives an overview of all the datasets. For simulation details and rationality analysis, please refer to the appendix.

4.3 Challenge.

Based on LiDAR and sparse IMUs, LIPD proposes a novel light-weight sensor setting for mocap in large-scale scenes. However, there are two main challenges, need to be solved, for reconstructing accurate human motions. The first is extracting valuable features from a sparsity-varying LiDAR point cloud, where the person in more than 20m away only contains a few points, and the second is finding an effective sensor-fusion method to make LiDAR and IMUs actually complement each other. We explore potential solutions for these problems and propose LIP as the baseline for this task.

5 Experiments

In this section, we compare our method with State-Of-The-Art (SOTA) methods qualitatively and quantitatively to show the superiority of the proposed LIP in Sect. 5.1. In addition, in Sect. 5.2, sufficient ablation studies are conducted to further demonstrate the rationality of the strategy for fusing multi-modal data and our architecture design. Moreover, we test multiple input configurations to prove the single LiDAR with four IMU sensors is an appropriate setting. For evaluation metrics, we use 1) **MPJPE**(*mm*): mean per root-relative joint position error; 2) **Mesh Err**(*mm*): mean per SMPL mesh vertex position error; 3) **Ang Err**(*deg*): mean per global joint rotation error to evaluate local pose, and 4) **CD**(*cm*): chamfer distance between vertices of SMPL mesh result and raw point cloud to evaluate global translation. For all these metrics, **lower means better**.

5.1 Comparison

We conduct comparisons to illustrate that our proposed LIP method enables more accurate capture for challenging motions and more precise translations in large-scale spaces. We select current SOTA TransPose [74] and LiDARCap [32] for comparative analysis. We use the model provided by TransPose [74] and reproduced LiDARCap [32] network (no released model) for comparison. Note that we test our reproduced model on the setting of LiDARCap [32] and achieves comparable performance as its paper claims. Thus, the comparison is fair. The superior results of LIP with different evaluation metrics are illustrated in Table 2. The visualization evaluation for mocap large-scale scenes is shown in Fig. 4. Benefiting from LiDAR-IMU hybrid modal input and our coarse-to-fine design, our method outperforms others by an obvious margin. To take advantage of the 3D distinguish-ability of the LiDAR point clouds, we design the *Pose-guided Translation Correction*. From the global view, LIP provides more accurate localization and more nature motion sequences. For the pose reconstruction of each frame, our result is aligned well with the point cloud. Fig. 5 further demonstrate the visualization results of local pose estimation on the testing data for more detailed comparison. Our method is obviously superior to others with results more close to the ground truth, especially for the challenging motions in LIPD. It is worth noting that LIP is still robust for some extreme cases with severe self-occlusions and external occlusions, like cases in the last row of Fig. 5 shows. In addition, we show more fancy results of our method for mocap in large-scale scenes in Fig. 6 to verify the generalization capability of LIP.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.

Table 2: Quantitative comparisons between LIP and related methods on our evaluation dataset. Note that LiDARHuman26M [32] and LIPD dataset only contains data with rate of 10fps, which is not applicable(N/A) for TransPose [74]. Also, the CD metric is not used for DIP-IMU [19] dataset since no translation is provided.

	TotalCapture [60]				DIP-IMU [19]			LiDARHuman26M [32]				LIPD(Ours)			
	MPJPE	Mesh Err	Ang Err	CD	MPJPE	Mesh Err	Ang Err	MPJPE	Mesh Err	Ang Err	CD	MPJPE	Mesh Err	Ang Err	CD
LiDARCap [32]	45.0	56.6	8.5	6.4	41.5	54.9	12.4	78.7	94.8	21.0	8.1	69.4	85.5	12.5	7.9
TransPose [74]	55.7	63.6	12.3	102.7	49.0	58.3	7.6	N/A	N/A	N/A	N/A	76.8	87.1	17.3	736.9
LIP	**30.0**	**39.5**	**7.4**	**0.7**	**30.2**	**39.1**	**9.6**	**60.7**	**71.6**	**11.6**	**3.5**	**48.9**	**59.8**	**11.3**	**3.2**

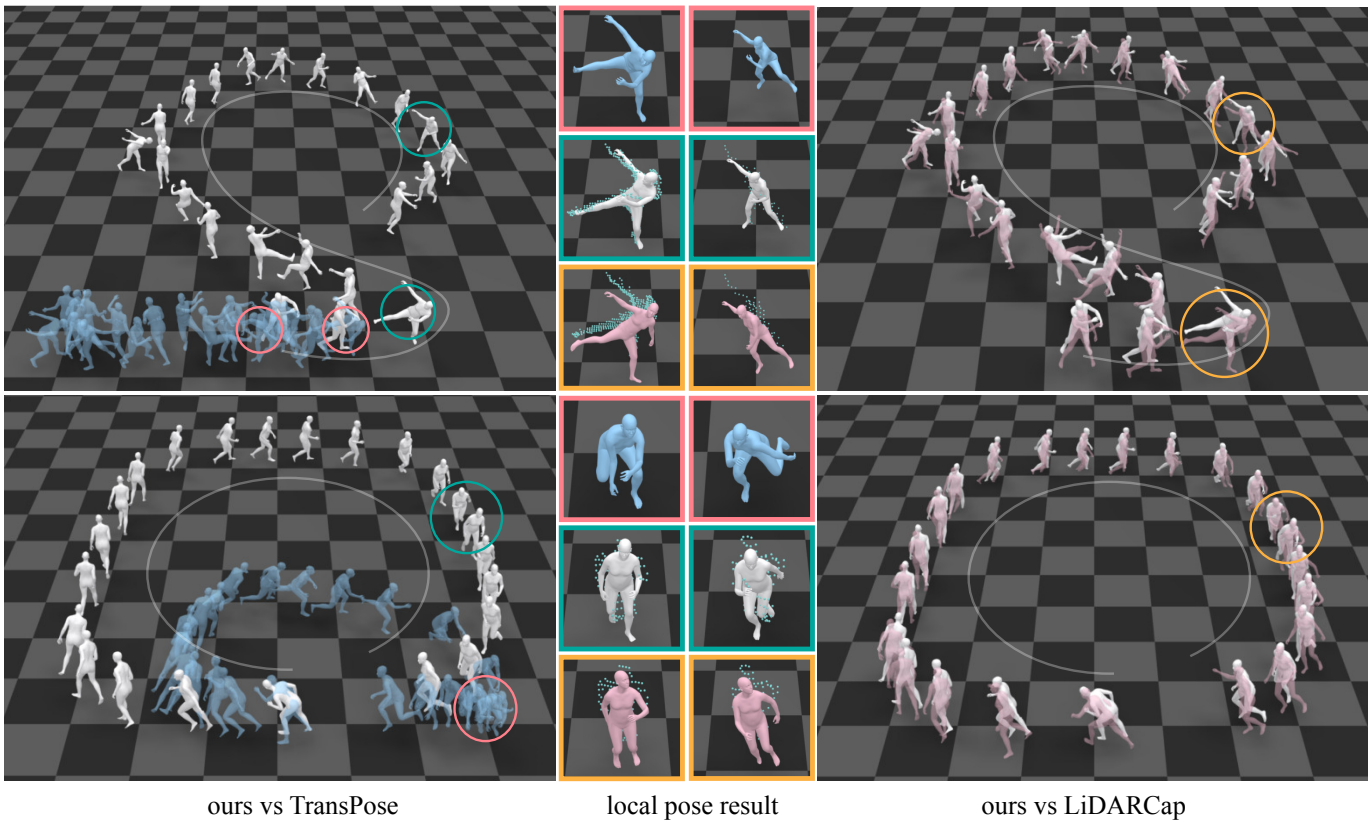


Fig. 4: Qualitative results of LIP (white) compared with TransPose [74] (blue) and LiDARCap [32] (pink). The trajectory is sketched by a grey curve. Detailed comparisons of some key frames are circled and zoomed in to show the local pose and alignment with point cloud. For intuitive visualization, we use the arithmetic mean of raw point cloud as the global translation for LiDARCap [32] in which translation inference is not included.

Table 3: Ablation study of different fusion strategies and the *Joint-map Estimator* (JE) block in our network.

Local pose	TotalCapture [60]			DIP-IMU [19]			LiDARHuman26M [32]			LIPD		
	MPJPE	Mesh Err	Ang Err	MPJPE	Mesh Err	Ang Err	MPJPE	Mesh Err	Ang Err	MPJPE	Mesh Err	Ang Err
FM-1	43.2	55.1	12.9	43.6	54.9	15.4	**59.4**	72.7	17.3	49.5	61.4	13.2
FM-2	32.3	42.3	7.8	31.2	41.3	9.7	61.8	74.3	14.2	50.9	62.3	11.3
FM-3	47.8	62.3	8.1	44.6	58.3	11.4	67.6	83.4	14.2	54.9	67.3	11.5
FM-4	36.7	42.4	**6.7**	37.3	49.0	10.0	59.6	**70.8**	12.6	52.4	64.5	**11.0**
Ours	**30.0**	**39.5**	7.4	**30.2**	**39.1**	**9.6**	60.7	71.6	**11.6**	**48.9**	**59.8**	11.3
Ours w/o JE	35.5	45.5	8.6	34.3	43.3	10.2	61.2	72.5	11.6	51.4	62.8	11.3

5.2 Ablation Study

We validate the effectiveness of our sensor-fusion method, network design, and the reasonability of the configuration of our multi-modal hybrid input by ablation studies.

Ablation Study on Network Designs. Based on the LiDAR-IMU multi-sensor configuration, it is important to make full use the information captured by both sensors and design an effective fusion method to make them complement each other. Thus, we first demonstrate the superiority of our LiDAR-IMU sensor-fusion method and compare with four other fusion strategies: **FM-1**) In the data processing phase, we directly merge the point cloud data and IMU data, which is a data-level fusion strategy; **FM-2**) We use GRU to extract high-dimensional features of IMU data and integrate it with the skeleton joints and root orientation estimated in the first stage from point cloud data; **FM-3**) The temporal features obtained from the raw point cloud are directly combined with IMU data as the input of the second stage; **FM-4**) We extracts features of IMU data and point cloud data respectively, and combine different modal features in high-dimensional feature space,

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.

Fig. 5: Qualitative comparison of capture performance for local details. Our method outperforms state-of-the-art works by more accurate local pose estimation on three different synthetic datasets and LIPD. LIPD-OC means the data with severe occlusions in LIPD. Note that LiDARHuman26M [32] is not applicable for TransPose [74] since it only provides 10fps mocap data, while frame rate of 60fps is required by TransPose [74].

Table 4: Ablation study of different global translation methods and our pose information guided processing (PG) in LIP. We demonstrate the evaluation results using CD metric.

Global translation	TotalCapture [60]	LiDARHuman26M [32]	LIPD
Average estimation	6.4	8.1	7.9
ICP estimation	2.0	4.7	4.1
Ours w/o PG	0.9	3.7	3.4
Ours	**0.7**	**3.5**	**3.2**

and then regress the pose parameters. We compare the above experiments on the testing set mentioned in Table 3. It illustrates that direct data-level or feature-level fusion strategies do not work well due to the large domain gap between these two modal data. Our coarse-to-fine fusion strategy achieves SOTA performance by regressing the coarse skeleton joints and root orientation as the intermediate bridge to align two modalities.

The second refinement stage of our network further benefits more accurate results in the estimation of joints and vertices. To verify the effectiveness of *Pose-guided Translation Correction* module, we conduct comparison with other location estimation methods and ablation studies. As Table 4 shows, the design of deviation regression and the usage of pose features benefit the performance a lot. To evaluate the generalization capability of LIP to point clouds captured in various ranges with different sparsity, we simulate a bunch of point clouds from TotalCapture [60] dataset at multiple virtual capture distances from 6 to 25 meters. Fig. 8 illustrates that our LIP performs consistently well in a wide range of about 20 meters.

Ablation Study on Input Configuration. To verify that single LiDAR with 4 IMUs is a necessary and compact configuration for high-quality mocap, we experiment various combinations of this two modal inputs with the same network architecture and training strategy. Table 5 gives quantitative comparisons which demonstrate that LiDAR-IMU hybrid input overall outperforms single modality input. Specifically, direct measurements of 3D scene provided by LiDAR reduce the ambiguities of regression from sparsely local inertia to 3D joint positions, which so that decrease the MPJPE and mesh error by more than 30 and 40 millimeters and angular error by 2 degrees over purely inertial input in average. Meanwhile, with the aid of local inertia, inaccurate estimation of joint positions and rotations on point cloud can be corrected in detail. Especially for LIPD datasets, of which the

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.

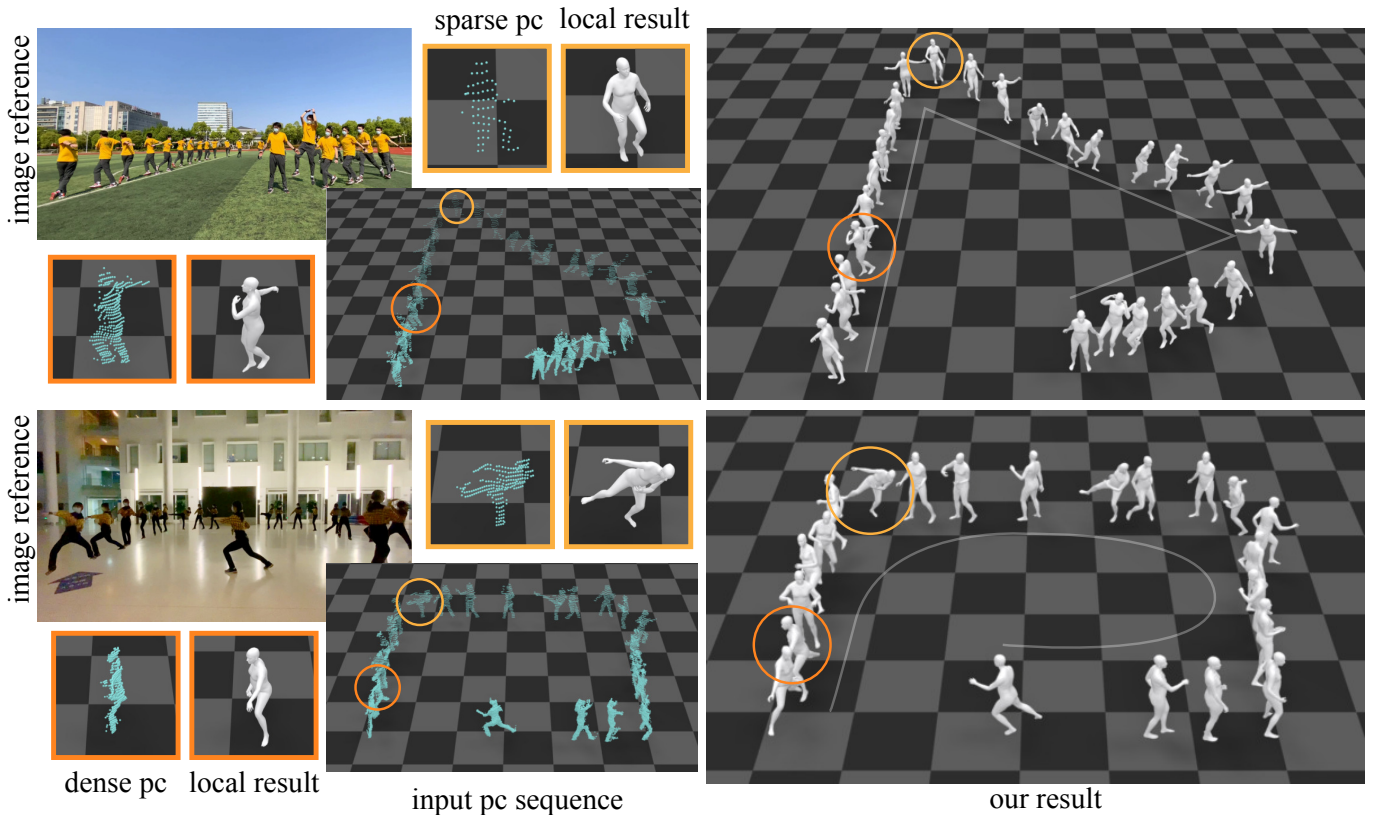


Fig. 6: Our results. We provide image reference, point cloud (pc) input, and mocap results in 3D scene in this figure. Some key frames are circled and zoomed in to show detailed local poses.

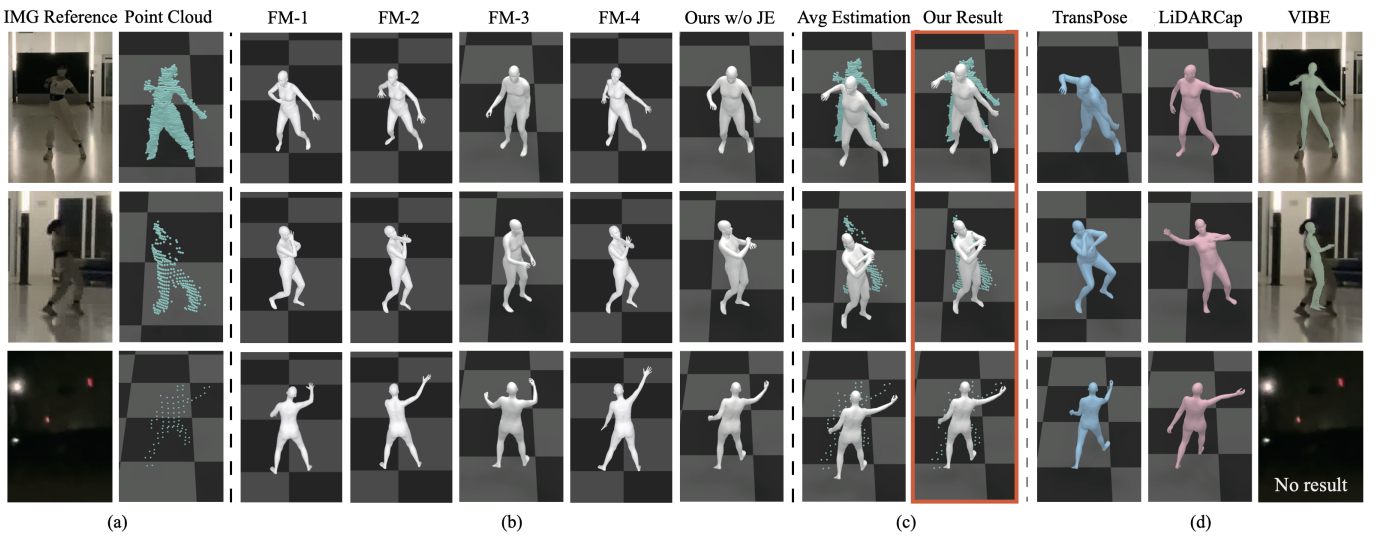


Fig. 7: Qualitative evaluations of our network design. We provide (a) image reference and point cloud input; (b) local pose evaluation; (c) global translation evaluation; (d) results of other methods.

point cloud data is real and imperfect with random noise, LIP improves the performance of LiDAR-only input by about 10 millimeters lower MPJPE, 15 millimeters lower mesh error and 3 degrees lower angular error. In addition, it is necessary to specify that the LiDAR-only performance comes from the same pipeline as LiDAR+n IMUs, except that no IMU data is fused with point cloud information.Different from LiDARCap [32], which is a one-stage method, the Module-I *Multi-modal Pose Estimation* in our pipeline estimates local body motion in two-stage. Besides, Table 5 illustrates that 4-IMU-aid configuration brings considerable improvements with only two more sensors compared with 2-IMU-aid version, and keeps acceptable performance gap to even 12-IMU-aid setting. Therefore, 4 IMU sensors is an appropriate choice to provide sufficient inertial aid while maintaining a light-weight capture setting, considering the complexity of capture system and convenience for performers.

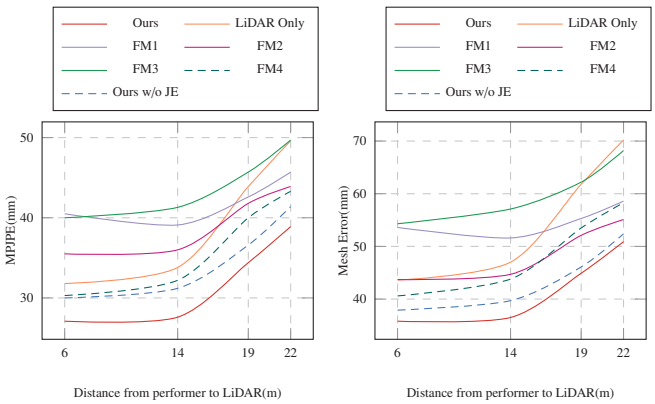
5.3 Discussion

Due to the low frame rate of the utilized LiDAR (10fps), it leads to unsmooth results when capturing extremely fast motions. We plan to improve the prediction modules using or Transformer to interpo-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.

Table 5: Ablation study for input configuration. We evaluate our choice of single LiDAR with 4 IMUs by experimenting the local motion estimation accuracy of various combinations of input modalities.

	TotalCapture [60]			DIP-IMU [19]			Lidarhuman26M [32]			LIPD		
	MPJPE	Mesh Err	Ang Err	MPJPE	Mesh Err	Ang Err	MPJPE	Mesh Err	Ang Err	MPJPE	Mesh Err	Ang Err
LiDAR Only	35.5	47.9	12.33	35.1	47.0	15.41	68.8	87.1	20.20	60.1	76.1	14.49
4 IMUs Only	76.3	93.2	10.65	91.4	103.1	14.04	90.0	106.6	14.14	69.1	83.8	12.77
5 IMUs only	59.9	74.0	9.59	79.0	88.8	11.96	71.9	87.4	12.86	50.3	60.7	11.65
LiDAR+2 IMUs	31.1	41.0	9.38	31.1	41.1	11.44	63.1	76.8	14.53	53.9	67.1	12.64
LiDAR+4 IMUs(Ours)	30.0	39.5	7.4	30.2	39.1	9.6	60.7	71.6	11.6	48.9	59.8	11.3
LiDAR+5 IMUs	30.1	39.7	7.4	30.7	39.9	9.7	57.9	69.0	11.2	44.9	55.4	10.6
LiDAR+12 IMUs	23.6	32.8	3.63	20.8	28.4	5.76	49.0	59.3	7.45	25.1	32.1	8.4



(a) (b)

Fig. 8: Evaluation for the generalization capability on various distances on MPJPE and Mesh Err.

Fig. 9: Applications of LIP. It enables conveniently capturing large-scale human performances and driving avatars in various virtual scenes.

late global poses, so as to recover the high-frequency motion details. Moreover, even though we have demonstrated that our method can generalize to different types of LiDAR(e.g. RS-LiDAR-M1 [53], OUSTER-0 [45] and OUSTER-1 [45]) and IMU sensor(e.g. Xsens Dot [68]) and Noitom [43]) in this paper, it's still interesting to enhance the capability of our method on handling the domain gaps between more different LiDAR sensors with various point distributions and perception ranges, which relies on further capturing a much larger dataset. It's also promising to extend our pipeline to multi-person or human-object interaction scenes.

6 Conclusion

In this paper, we propose a new solution, LiDAR-aid Inertial Poser(LIP), for conveniently capturing human motions in large-scale scenarios, which is occlusion-free and environment-independent. As shown in Fig. 9, the recovered human poses and global translations further enable various mix-reality applications like driving avatars in virtual scenes. To explore the new multi-modal setting (light-weight 4 IMUs plus one LiDAR), we design an effective fusion strategy for taking advantage of both LiDAR point cloud and IMU measurements. We also eliminate the translation deviation in a pose-guided manner and gain accurate global trajectories and natural consecutive actions. Importantly, we propose a huge LiDAR-IMU hybrid mocap dataset for future research of human motion analysis. Extensive experimental results demonstrate the effectiveness of our method and module designs, for convenient and high-quality human motion capture in a large scene. We believe our approach and dataset serve as a critical step for light-weight and global human mocap in large-scale scenes, with many potential applications for VR/AR, gaming, filming, or entertainment.

Acknowledgments

This work was supported by NSFC (No.62206173), Shanghai Sailing Program (No.22YF1428700), and Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

References

[1] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3D human pose estimation. In *British Machine Vision Conference (BMVC)*, 2009. doi: 10.5244/C.27.45

[2] S. Andrews, I. Huerta, T. Komura, L. Sigal, and K. Mitchell. Real-time physics-based motion capture with sparse sensors. In *Proceedings of the 13th European conference on visual media production (CVMP 2016)*, pp. 1–10, 2016.

[3] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *International Conference on Computer Vision (ICCV)*, 2011. doi: 10.1109/ICCV.2011.6126356

[4] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In B. Leibe, J. Matas, N. Sebe, and M. Welling, eds., *Computer Vision – ECCV 2016*, pp. 561–578. Springer International Publishing, Cham, 2016.

[5] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Computer Vision and Pattern Recognition (CVPR)*, 1998. doi: 10.1109/CVPR.1998.698581

[6] M. Burenius, J. Sullivan, and S. Carlsson. 3D pictorial structures for multiple view articulated pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2013. doi: 10.1109/CVPR.2013.464

[7] P. Cong, X. Zhu, F. Qiao, Y. Ren, X. Peng, Y. Hou, L. Xu, R. Yang, D. Manocha, and Y. Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19608–19617, 2022.

[8] Y. Dai, Y. Lin, C. Wen, S. Shen, L. Xu, J. Yu, Y. Ma, and C. Wang. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *Proceedings of the IEEE/CVF*

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.

Conference on Computer Vision and Pattern Recognition, pp. 6792–6802, 2022.

[9] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *ACM SIGGRAPH 2008 papers*, pp. 1–10. 2008.

[10] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] A. Gilbert, M. Trumble, C. Malleson, A. Hilton, and J. Collomosse. Fusing visual and inertial sensors with semantics for 3d human pose estimation. *International Journal of Computer Vision*, 127(4):381–397, 2019.

[12] K. Guo, J. Taylor, S. Fanello, A. Tagliasacchi, M. Dou, P. Davidson, A. Kowdle, and S. Izadi. Twinfusion: High framerate non-rigid fusion through fast correspondence tracking. In *International Conference on 3D Vision (3DV)*, pp. 596–605, 2018.

[13] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 38(2):14:1–14:17, 2019.

[14] M. Habermann, W. Xu, M. Zollhofer, G. Pons-Moll, and C. Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] Y. He, A. Pang, X. Chen, H. Liang, M. Wu, Y. Ma, and L. Xu. Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11400–11411, 2021.

[16] T. Helten, M. Muller, H.-P. Seidel, and C. Theobalt. Real-time body tracking with one depth camera and inertial sensors. In *Proceedings of the IEEE international conference on computer vision*, pp. 1105–1112, 2013.

[17] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *Journal of Selected Topics in Signal Processing*, 6(5):538–552, 2012. doi: 10.1109/JSTSP.2012.2196975

[18] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 International Conference on 3D Vision (3DV)*, pp. 421–430, 2017. doi: 10.1109/3DV.2017.00055

[19] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018.

[20] C. Ionescu, D. Papava, V. Olaru, and C. 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.

[21] J. Jiang, P. Streli, H. Qiu, A. Fender, L. Laich, P. Snape, and C. Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European Conference on Computer Vision*, pp. 443–460. Springer, 2022.

[22] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *International Conference on Computer Vision (ICCV)*, 2015. doi: 10.1109/ICCV.2015.381

[23] T. Kaichi, T. Maruyama, M. Tada, and H. Saito. Resolving position ambiguity of imu-based human pose with a single rgb camera. *Sensors*, 20(19):5453, 2020.

[24] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[25] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[26] W. Kim, M. S. Ramanagopal, C. Barto, M.-Y. Yu, K. Rosaen, N. Goumas, R. Vasudevan, and M. 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.

[27] M. Kocabas, N. Athanasiou, and M. 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 (CVPR)*, June 2020.

[28] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11127–11137, October 2021.

[29] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11605–11614, 2021.

[31] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6050–6059, 2017.

[32] J. Li, J. Zhang, Z. Wang, S. Shen, C. Wen, Y. Ma, L. Xu, J. Yu, and C. 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*, pp. 20502–20512, 2022.

[33] R. Li, S. Yang, D. A. Ross, and A. Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13401–13412, 2021.

[34] H. Liang, Y. He, C. Zhao, M. Li, J. Wang, J. Yu, and L. Xu. Hybridcap: Inertia-aid monocular capture of challenging human motions. *arXiv preprint arXiv:2203.09287*, 2022.

[35] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015.

[36] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[37] Z. Luo, R. Hachiuma, Y. Yuan, and K. Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34, 2021.

[38] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[39] C. Malleson, J. Collomosse, and A. Hilton. Real-time multi-person motion capture from multi-view video and imus. *International Journal of Computer Vision*, pp. 1–18, 2019.

[40] C. Malleson, J. Collomosse, and A. Hilton. Real-time multi-person motion capture from multi-view video and imus. *International Journal of Computer Vision*, 128(6):1594–1611, 2020.

[41] C. Malleson, A. Gilbert, M. Trumble, J. Collomosse, A. Hilton, and M. Volino. Real-time full-body motion capture from video and imus. In *2017 International Conference on 3D Vision (3DV)*, pp. 449–457. IEEE, 2017.

[42] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pp. 506–516. IEEE, 2017.

[43] Noitom Motion Capture Systems. `https://www.noitom.com/`, 2015.

[44] OptiTrack Motion Capture Systems. `https://www.optitrack.com/`, 2009.

[45] OUSTER High Performance Digital Lidar Solutions. `https://ouster.com/`, 2021.

[46] A. K. Patil, A. Balasubramanyam, J. Y. Ryu, P. K. BN, B. Chakravarthi, and Y. H. Chai. Fusion of multiple lidars and inertial sensors for the real-time pose tracking of human motion. *Sensors*, 20(18):5342, 2020.

[47] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[48] G. Pons-Moll, A. Baak, T. Helten, M. Müller, H.-P. Seidel, and B. Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 663–670. IEEE, 2010.

[49] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

[50] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11488–11499, October 2021.

[51] H. Rhodin, N. Robertini, C. Richardt, H.-P. Seidel, and C. Theobalt. A

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.

versatile scene model with differentiable visibility applied to generative pose estimation. In *International Conference on Computer Vision (ICCV)*, 2015. doi: 10.1109/ICCV.2015.94

[52] N. Robertini, D. Casas, H. Rhodin, H.-P. Seidel, and C. Theobalt. Model-based outdoor performance capture. In *International Conference on 3D Vision (3DV)*, 2016.

[53] RS-LiDAR-M1 Leading the large-scale production of automotive-grade MEMS LiDAR. `https://www.robosense.cn/en/rslidar/RS-LiDAR-M1`, 2019.

[54] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[55] L. Sigal, A. O. Bălan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 2010. doi: 10.1007/s11263-009-0273-6

[56] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision (IJCV)*, 98(1):15–48, 2012. doi: 10.1007/s11263-011-0493-4

[57] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[58] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of Gaussians body model. In *International Conference on Computer Vision (ICCV)*, 2011.

[59] C. Theobalt, E. de Aguiar, C. Stoll, H.-P. Seidel, and S. Thrun. Performance capture from multi-view video. In *Image and Geometry Processing for 3-D Cinematography*, pp. 127–149. Springer, 2010.

[60] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pp. 1–13. University of Surrey, 2017.

[61] Vicon Motion Capture Systems. `https://www.vicon.com/`, 2010.

[62] D. Vlasic, R. Adelsberger, G. Vannucci, J. Barnwell, M. Gross, W. Matusik, and J. Popović. Practical motion capture in everyday surroundings. *ACM transactions on graphics (TOG)*, 26(3):35–es, 2007.

[63] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. 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)*, pp. 601–617, 2018.

[64] T. Von Marcard, G. Pons-Moll, and B. Rosenhahn. Human pose estimation from video and imus. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1533–1547, 2016.

[65] T. Von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, vol. 36, pp. 349–360. Wiley Online Library, 2017.

[66] X. Wei, P. Zhang, and J. Chai. Accurate realtime full-body motion capture using a single depth camera. *SIGGRAPH Asia*, 31(6):188:1–12, 2012.

[67] A. Winkler, J. Won, and Y. Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–8, 2022.

[68] Xsens Technologies B.V. `https://www.xsens.com/`, 2011.

[69] L. Xu, Y. Liu, W. Cheng, K. Guo, G. Zhou, Q. Dai, and L. Fang. Flycap: Markerless motion capture using multiple autonomous flying cameras. *IEEE Transactions on Visualization and Computer Graphics*, 24(8):2284–2297, Aug 2018.

[70] L. Xu, Z. Su, L. Han, T. Yu, Y. Liu, and L. FANG. Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercialrgbd cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[71] L. Xu, W. Xu, V. Golyanik, M. Habermann, L. Fang, and C. Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[72] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 37(2):27:1–27:15, 2018.

[73] X. Yi, Y. Zhou, M. Habermann, S. Shimada, V. Golyanik, C. Theobalt, and F. Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[74] X. Yi, Y. Zhou, and F. Xu. Transpose: real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021.

[75] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[76] Y. Yuan and K. Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. *Advances in Neural Information Processing Systems*, 33:21763–21774, 2020.

[77] A. Zanfir, E. G. Bazavan, M. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu. Neural descent for visual 3d human pose and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14484–14493, 2021.

[78] Z. Zhang, C. Wang, W. Qin, and W. Zeng. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2200–2209, 2020.

[79] Z. Zheng, T. Yu, H. Li, K. Guo, Q. Dai, L. Fang, and Y. Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *European Conference on Computer Vision (ECCV)*, Sept 2018.

[80] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019.

[81] X. Zhu, Y. Ma, T. Wang, Y. Xu, J. Shi, and D. Lin. Ssn: Shape signature networks for multi-class object detection from point clouds. In *European Conference on Computer Vision*, pp. 581–597. Springer, 2020.

[82] X. Zhu, H. Zhou, T. Wang, F. Hong, W. Li, Y. Ma, H. Li, R. Yang, and D. Lin. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[83] J. Ziegler, H. Kretzschmar, C. Stachniss, G. Grisetti, and W. Burgard. Accurate human motion capture in large areas by combining imu-and laser-based people tracking. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 86–91. IEEE, 2011.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on August 16,2026 at 10:30:06 UTC from IEEE Xplore. Restrictions apply.