LiveHPS: LiDAR-based Scene-level Human Pose and Shape Estimation in Free Environment

Yiming Ren[1], Xiao Han[1], Chengfeng Zhao[1], Jingya Wang[1], Lan Xu[1], Jingyi Yu[1], Yuexin Ma[1,*]
[1] ShanghaiTech University
{renym2022,mayuexin}@shanghaitech.edu.cn

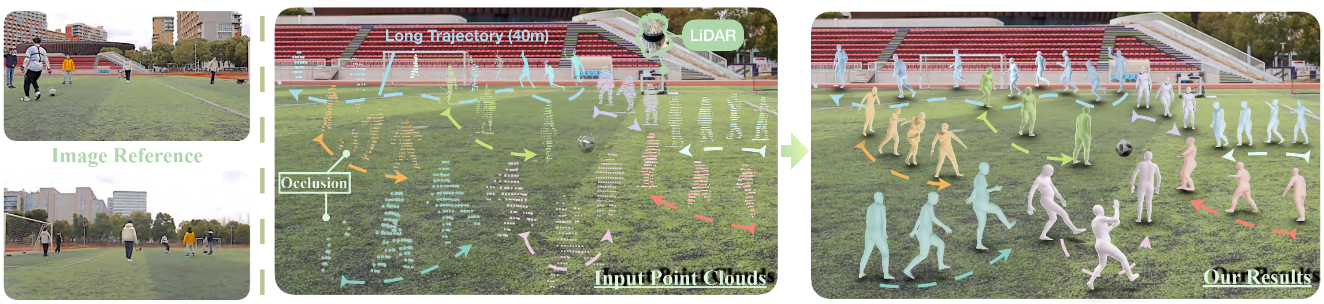

Figure 1. We propose a novel single-LiDAR-based approach for 3D HPS in large-scale scenarios, which is not limited to fixed studios, light conditions, and wearable devices. Our method predicts full human SMPL parameters(pose, shape, translation) from consecutive LiDAR point clouds and performs well for challenging poses and occlusion situations.

Abstract

*For human-centric large-scale scenes, fine-grained modeling for 3D human global pose and shape is significant for scene understanding and can benefit many real-world applications. In this paper, we present **LiveHPS**, a novel single-**Li**DAR-based approach for scene-le**ve**l **H**uman **P**ose and **S**hape estimation without any limitation of light conditions and wearable devices. In particular, we design a distillation mechanism to mitigate the distribution-varying effect of LiDAR point clouds and exploit the temporal-spatial geometric and dynamic information existing in consecutive frames to solve the occlusion and noise disturbance. LiveHPS, with its efficient configuration and high-quality output, is well-suited for real-world applications. Moreover, we propose a huge human motion dataset, named **FreeMotion**, which is collected in various scenarios with diverse human poses, shapes and translations. It consists of multi-modal and multi-view acquisition data from calibrated and synchronized LiDARs, cameras, and IMUs. Extensive experiments on our new dataset and other public datasets demonstrate the SOTA performance and robustness of our approach.*

*Corresponding author. This work was supported by NSFC (No.62206173), Natural Science Foundation of Shanghai (No.22dz1201900), Shanghai Sailing Program (No.22YF1428700), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

1. Introduction

Human pose and shape estimation (HPS) is aimed at reconstructing 3D digital representations of human bodies, such as SMPL [32], using data captured by sensors. It is significant for two primary applications: one in motion capture for the entertainment industry, including film, augmented reality, virtual reality, mixed reality, etc.; and the other in behavior understanding for the robotics industry, covering domains like social robotics, assistive robotics, autonomous driving, human-robot interaction, and beyond.

While optical-based methods [15, 16, 26, 28, 41] have seen significant advancements in recent years, their efficacy is limited due to the camera sensor's inherent sensitivity to variations in lighting conditions, rendering them impractical for use in uncontrolled environments. In contrast, inertial methods [37, 55, 60, 61] utilize body-mounted inertial measurement units (IMUs) to derive 3D poses, which is independent of lighting and occlusions. However, these methods necessitate the use of wearable devices, struggle with drift issues over time, and fail to capture human body shapes and precise global translations.

LiDAR is a commonly used perception sensor for robots and autonomous vehicles [9, 33, 65, 66] due to its accurate depth sensing without light interference. Recent advances [42] in HPS are turning to utilize LiDAR for capturing high-quality SMPLs in the wild. LiDARCap [30] proposes a GRU-based approach for estimating only human pose parameters from LiDAR point clouds. MOVIN [20] uses a CVAE framework to link point clouds with human poses for both human pose and global translation estimation. Nevertheless, these approaches lack the capability to estimate body shapes, and further, they disregard the challenging characteristics of LiDAR point clouds, leading to an unstable performance in real-world scenarios. First, the distribution and pattern of LiDAR point clouds vary across different capture distances and devices. Second, the view-dependent nature of LiDAR results in incomplete point clouds of the human body, affected by self-occlusion or external obstruction. Third, real-captured LiDAR point clouds invariably contain noise in complex scenarios, caused by the reflection interference or carry-on objects. These properties all bring challenges for accurate and robust HPS in extensive, uncontrolled environments.

Considering above intractable problems of LiDAR point cloud, we introduce LiveHPS, a novel single-LiDAR-based approach for capturing high-quality human pose, shape, and global translation in large-scale free environment, as shown in Fig. 1. The deployment-friendly single-LiDAR setting is unrestricted in acquisition sites, light conditions, and wearable devices, which can benefit many practical applications. In order to improve the robustness for tackling point distribution variations, we design an **Adaptive Vertex-guided Distillation** module to make diverse point distributions align with the regular SMPL mesh vertex distribution in high-level feature space by a prior consistency loss. Moreover, to reduce the influence of occlusion and noise, we propose a **Consecutive Pose Optimizer** to explore the geometric and dynamic information existing in temporal and spatial spaces for pose refinement by attention-based feature enhancement. In addition, a **Skeleton-aware Translation Solver** is also presented to eliminate the effect of incomplete and noised point cloud on accurate estimation for human global translation. In particular, we introduce the scene-level unidirectional Chamfer distance (SUCD) from the input point cloud to the estimated human mesh vertex in global coordinate system as a new evaluation measurement for LiDAR-based HPS, which can reflect the fine-grained geometry error and translation error between the prediction and the ground truth.

It is worth noting that we also introduce **FreeMotion**, a novel huge motion dataset captured in diverse large-scale real scenarios with multiple persons, which contains multi-modal data(LiDAR point clouds, RGB image and IMUs), multi-view data(front, back and side), and comprehensive SMPL parameters(pose, shape and global translation). Through extensive experiments and ablation studies on FreeMotion and other public datasets, our method outperforms others by a large margin.

Our main contributions can be summarized as follows:

- We present a novel single-LiDAR-based method for 3D HPS in large-scale free environment, which achieves state-of-the-art performance.
- We propose an effective vertex-guided adaptive distillation module, consecutive pose optimizer, and skeleton-aware translation solver to deal with the distribution-varied, incomplete, and noised LiDAR point clouds.
- We present a new motion dataset captured in diverse real scenarios with rich modalities and annotations, which can facilitate further research of in-the-wild HPS.

2. Related Work

2.1. Optical-based Methods

Optical motion capture technology has advanced from initial marker-based systems [38, 49, 50] that rely on camera-tracked markers to reconstruct a 3D mesh, to markerless systems [1, 5, 12, 21, 34, 39, 43, 44, 47]. Despite they can get high-accuracy results, these systems are often expensive and require elaborate setup and calibration. To mitigate these challenges, monocular mocap methods using optimization [4, 17, 27, 29] and regression [22, 23, 25, 64], along with template-based, probabilistic [14–16, 56, 57], and semantic-modeling techniques [26], have emerged to address monocular system limitations. Nonetheless, these approaches still suffer from inherent light sensitivity and depth ambiguity. Some strategies[2, 13, 45, 54, 63] use depth cameras, yet these cameras have a limited sensing range and are ineffective in outdoor scenes.

2.2. Inertial-based Methods

Unlike optical systems, inertial motion capture systems [55] are not affected by light conditions and occlusions. They generally need numerous IMUs attached to form-fitting suits, a setup that can be heavy and inconvenient, motivating interest in more sparse configurations, such as six-IMU setting [18, 51, 60, 61] and four-IMU setting [42]. However, these methods suffer from drift errors over time, cannot provide precise shape and global translation, and require wearable devices, not practical for daily-life scenarios.

2.3. LiDAR-based Methods

With precise long-range depth-sensing ability, LiDAR has emerged as a key sensor in robotics and autonomous vehicles [8, 40, 62, 65, 66]. LiDAR can provide precise depth information and global translation in expansive environments, remaining uninfluenced by lighting conditions, enabling robust 3D HPS. Recently, PointHPS [7] provides a cascaded network architecture for pose and shape estimation from point clouds. However, it is applicable for dense point clouds rather than sparse LiDAR point clouds. LiDARCap [30] employs a graph-based convolutional network to predict daily human poses in LiDAR-captured large-scale scenes. MOVIN [20] presents a generative method for estimating both pose and global translation. However, these methods cannot predict full SMPL parameters (pose, shape, and global translation) and are fragile for complex real scenarios with occlusion and noise.

2.4. 3D Human Motion Datasets

Data-driven 3D HPS have gained traction in recent years benefiting from extensive labeled datasets. Indoor marker-based datasets [19, 46] use multi-view camera systems to record daily motions. AMASS [35] unifies these datasets, providing a standardized benchmark for network training. Marker-less datasets such as MPI-INF-3DHP [36] and AIST++ [31] capture more complex poses without constraint of the wearable devices, all above datasets are still confined to indoor settings. Outdoor motion capture datasets [24, 52] capture motions in the wild but lack accurate depth information, hindering scene-level human motion research. HuMMan [6] constitutes a mega-scale database that offers high-resolution scans for subjects, and MOVIN [20] provides motion data from multi-camera capture system with point clouds, but both datasets are limited in short-range scenes. [10, 11, 59] are proposed for human motion capture in large-scale scenes using environment-involved optimization, but they are limited in a single-person setting. Recently, LiDARHuman26M [30] and LIP [42] provide LiDAR-captured motion dataset in large scenes, but both datasets exclusively provide pose parameters of SMPL in single-person scenarios. In contrast, we propose a large-scale LiDAR-based motion dataset with full SMPL parameter annotations. It comprises challenging scenarios with occlusions and interactions among multiple persons and objects, which has great practical significance.

3. Methodology

We propose a single-LiDAR-based approach named LiveHPS for scene-level 3D human pose and shape estimation in large-scale free environments. The overview of our pipeline is shown in Fig. 2. We take consecutive 3D single-person point clouds as input and aim to acquire consistently accurate local pose, human shape, and global translation without any limitation of acquisition sites, light conditions, and wearable devices. There are three main procedures in our network, including point-based body tracker (Sec. 3.2), consecutive pose optimizer (Sec. 3.3), and attention-based multi-head SMPL solver (Sec. 3.4). First, we utilize the point-based body tracker to extract point-wise features and predict the human body joint positions. Second, we propose the attention-based temporal-spatial feature enhancement mechanism to acquire refined joint positions using joint-wise geometric and relationship features. Finally, we design an attention-based multi-head solver to regress the human SMPL parameters including human local pose, shape and global translation from the refined body skeleton.

3.1. Preliminaries

LiveHPS takes a consecutive sequence of single-person point clouds with T frames as input. As raw point clouds have various numbers of points at different times t, we implement normalization process by resampling each frame to a fixed $N_{fps} = 256$ points utilizing the farthest point sampling algorithm (FPS) and subtracting the average locations $\mathbf{loc}(t) \in \mathbb{R}^3$ of the raw data. $\mathbf{P}(t) \in \mathbb{R}^{3N_{fps}}$ denotes the pre-processed input at time t.

We define N_J as the number of body joints and N_V as the number of body vertices on SMPL mesh; $\hat{\mathbf{J}}(t)$, $\widetilde{\mathbf{J}}^{GT}(t) \in \mathbb{R}^{3N_J}$ as predicted and ground-truth root-relative joint positions at time t, repectively; $\hat{\mathbf{V}}(t), \widetilde{\mathbf{V}}^{GT}(t) \in \mathbb{R}^{3N_V}$ as predicted and ground-truth vertex positions. Our network prediction consists of $\hat{\theta}(t) \in \mathbb{R}^{6N_J}$, $\hat{\beta} \in \mathbb{R}^{10}$ and $\hat{Tr}(t) \in \mathbb{R}^3$, the pose, shape, and global translation parameters of SMPL. $\theta^{GT}(t)$, β^{GT} and $Tr^{GT}(t)$ are corresponding ground truth. We use 6D-rotation-based pose representation.

3.2. Point-based Body Tracker

For the input of our pre-processed consecutive point clouds, we extract the point-wise feature following the PointNet-GRU structure proposed by LIP [42] and regress the human body joint positions with an MLP decoder. Considering the irregular distribution of LiDAR point clouds vary across different capture distances and devices, and are also effected by occlusion and noise (Fig. 1), we design a **Vertex-guided Adaptive Distillation (VAD)** mechanism to unify the point distribution to facilitate the training of the network and improve the robustness. Because the vertices of SMPL mesh have relatively regular representation, we make diverse point distributions aligned with the mesh vertex distribution in high-level feature space by distillation, as Fig. 2 shows.

Firstly, we use the global translation $Tr^{GT}(t)$ to align the LiDAR point cloud $\mathbf{P}(t)$ with the ground truth mesh vertex $\widetilde{\mathbf{V}}^{GT}(t)$ and utilize k-Nearest-Neighbours (kNN) algorithm to sample the corresponding vertices, defined as $\widetilde{\mathbf{V}}^{GT}_{pc}(t)$. Then, we use $\widetilde{\mathbf{V}}^{GT}_{pc}(t)$ to pre-train a vertex body tracker to regress the joint positions $\hat{\mathbf{J}}_v(t)$. We use the mean squared error (MSE) loss $L_{mse}(\hat{\mathbf{J}}_v)$ for supervision:

$$\mathcal{L}_{mse}(\hat{\mathbf{J}}_v) = \sum_t \parallel \hat{\mathbf{J}}_v(t) - \widetilde{\mathbf{J}}^{GT}(t) \parallel_2^2 . \quad (1)$$

Subsequently, we input sequential point clouds $\mathbf{P}(t)$ and their corresponding vertex data $\widetilde{\mathbf{V}}^{GT}_{pc}(t)$ into two independent body trackers to obtain point-wise features $F_p(t) \in \mathbb{R}^k$

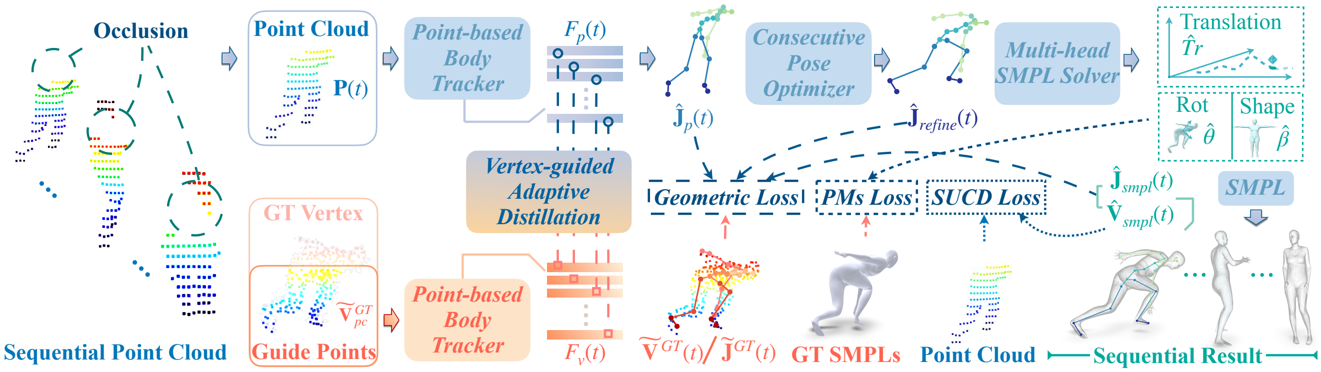

Figure 2. The pipeline of LiveHPS. With sequential LiDAR point clouds as input, LiveHPS consists of three critical modules to obtain human SMPL parameters, including a point-based body tracker to distill the pose-prior information, a consecutive pose optimizer to refine the pose via utilizing joint-wise features, and a multi-head SMPL solver to regress parameters of human models.

and $F_v(t) \in \mathbb{R}^k$, respectively, where $k = 1024$. Notably, two point-based body tracker networks share distinct weights and we freeze the pre-trained parameters of the vertex body tracker during training. To align real point distributions with the regular vertex distribution, we employ a pose-prior consistency loss $\mathcal{L}_{pc}$ to minimize the high-level feature distance between LiDAR point clouds and guided vertices. The distillation procedure enables our feature extractor to own the ability to maintain insensitivity under vastly differentiated data distribution. Finally, we leverage an MLP decoder to predict the joint positions $\hat{\mathbf{J}}_p$. A combined loss $\mathcal{L}_{prior}$ consisting of $\mathcal{L}_{mse}(\hat{\mathbf{J}}_p)$ and $\mathcal{L}_{pc}$ is utilized to train the network, which is formulated as below

$$\mathcal{L}_{mse}(\hat{\mathbf{J}}_p) = \sum_t \parallel \hat{\mathbf{J}}_p(t) - \widetilde{\mathbf{J}}^{GT}(t) \parallel_2^2, \tag{2}$$

$$\mathcal{L}_{pc} = \sum_t F_v(t) \log(\frac{F_v(t)}{F_p(t)}), \tag{3}$$

$$\mathcal{L}_{prior} = \lambda_1 \mathcal{L}_{mse}(\hat{\mathbf{J}}_p) + \lambda_2 \mathcal{L}_{PC}, \tag{4}$$

where λ_1 and λ_2 are hyper-parameters, and we set $\lambda_1 = 1$ and $\lambda_2 = 10^3$ in our experiments. During inference, the VAD process is not required.

3.3. Consecutive Pose Optimizer

We have already obtained the joint positions of human poses from the point-based body tracker. Considering that human motions are coherent at time sequence and different joints of the human body usually execute the action with relative dynamic constraints, we propose a **Consecutive Pose Optimizer (CPO)** (Fig. 3) to refine the body skeleton using consecutive joint-wise geometry features and relationship features in temporal and spatial spaces, which can further reduce the effect of incomplete and noised point clouds. Specifically, we utilize the concatenation of point-wise feature $F_p(t) \in \mathbb{R}^k$ and the predicted joint positions $\hat{\mathbf{J}}_p(t)$ as the initial joint-wise feature input. To capture the motion

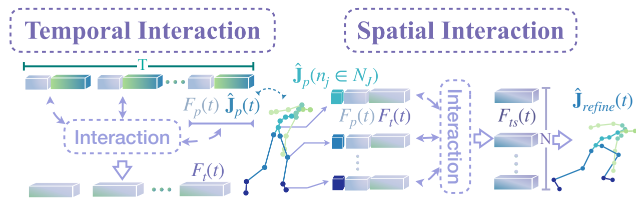

Figure 3. The detailed feature interaction mechanism in CPO. The same network architecture is applied in both consecutive pose optimizer and multi-head solver(pose and shape) except the decoder. Here we take the consecutive pose optimizer as the reference.

consistency in sequence, we use linear transformations to generate $Q(t)$, $K(t)$, and $V(t)$ in each frame and conduct temporal interaction to learn the motion-consistent feature $F_t(t) \in \mathbb{R}^{k_2}$ for each joint, where $k_2 = 256$. This temporal interaction process guides the estimation of more reasonable continuous human motions, especially for occluded situations. Then, we use the dynamic and geometric constraints among joints to further enhance the joint feature via spatial feature interaction. The input $F_j(n_j \in N_J) \in \mathbb{R}^{k+k_2+3}$ consists of the point-wise feature $F_p(t) \in \mathbb{R}^k$, temporal interaction feature $F_t(t) \in \mathbb{R}^{k_2}$, and each joint feature $\hat{\mathbf{J}}_p(n_j \in N_J) \in \hat{\mathbf{J}}_p(t)$. We generate $Q(n_j)$, $K(n_j)$ and $V(n_j)$ with linear mapping for each joint and conduct the spatial joint-to-joint interaction to get the enhanced feature $F_{ts}(t) \in \mathbb{R}^{k_3}$, where $k_3 = 512$. The feature interaction matrix can be formulated as:

$$\mathcal{F}_{interaction} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{5}$$

Finally, we regress the refined joint positions $\hat{\mathbf{J}}_{refine}(t)$ from the enhanced feature and the loss function is

$$\mathcal{L}_{mse}(\hat{\mathbf{J}}_{refine}) = \sum_t \parallel \hat{\mathbf{J}}_{refine}(t) - \widetilde{\mathbf{J}}^{GT}(t) \parallel_2^2. \tag{6}$$

3.4. Multi-head SMPL Solver

In the last stage, we propose an attention-based multi-head solver to regress the SMPL [32] parameters $\hat{\theta}(t)$, $\hat{\beta}$, $\hat{Tr}(t)$ from refined joint positions and the input point cloud. Because the pose and the shape reflect local geometry of human body, they can be determined by root-relative joint features obtained in last stage. We utilize the same network structure as CPO as the pose solver and shape solver to get $\hat{\theta}(t)$ and $\hat{\beta}$. However, the global translation could be obtained only from the root-relative local geometry features. Previous methods [30, 42] usually take the average position of the body point cloud as the global location or directly regress the global translation. However, due to the interference of occlusion and noise, their predicted results are unstable in consecutive frames. In contrast, we simplify the task of predicting global translation to predict the bias between the average position of point cloud and the real 3D location. Thus, we propose a **Skeleton-aware Translation Solver** underpinned by a cross-attention architecture, which intelligently integrates skeletal and original point cloud data to get more accurate translation estimation. We employ point cloud $\mathbf{P}(t)$ and refined root-relative joint positions $\hat{\mathbf{J}}_{refine}(t)$ as the input, utilizing the cross-attention to match the geometric information of joints with the point cloud. We generate the $Q(t)$ from refined joint positions and $K(t)$, $V(t)$ from point cloud. The feature interaction matrix can be formulated as Eq. 5. The decoder outputs the bias, which can be added to the average location $\mathbf{loc}(t)$ of raw point cloud to get the global translation $\hat{Tr}(t)$. Finally, we use SMPL model to generate the human skeleton joint positions and mesh vertex positions as below.

$$\hat{\mathbf{J}}_{smpl}(t), \hat{\mathbf{V}}_{smpl}(t) = \text{SMPL}(\hat{\theta}(t), \hat{\beta}, \hat{Tr}(t)). \quad (7)$$

The loss function for the multi-head solver is formulated as:

$$\begin{aligned} \mathcal{L}_{solver} =& \lambda_3 \mathcal{L}_{mse}(\hat{\mathbf{J}}_{smpl}) + \lambda_4 \mathcal{L}_{mse}(\hat{\mathbf{V}}_{smpl}) \\ &+ \lambda_5 \mathcal{L}_{mse}(\hat{\theta}(t)) + \lambda_6 \mathcal{L}_{mse}(\hat{\beta} \\ &+ \lambda_7 \mathcal{L}_{mse}(\hat{Tr}(t)) + \lambda_8 \mathcal{L}_{SUCD}, \end{aligned} \quad (8)$$

where λ_3, λ_4, λ_5, λ_6, λ_7 are hyper-parameters with $\lambda_3 = \frac{100}{N_j}$, $\lambda_4 = \frac{100}{N_v}$, $\lambda_5 = \frac{1}{5}$, $\lambda_6 = 1$, $\lambda_7 = 1$ and $\lambda_8 = 10^3$.

Because the raw point cloud contains the real pose, shape, and global translation information, it can be taken as an extra supervision which is ignored by previous methods. In particular, we introduce a novel scene-level unidirectional Chamfer distance (SUCD) loss by calculating the unidirectional Chamfer distance from the raw point cloud to the predicted mesh vertices. It provides a comprehensive evaluation for all predicted SMPL parameters, denoted as

$$\mathcal{L}_{SUCD} = \sum_t \frac{1}{|\mathbf{P}(t)|} \sum_{x \in \mathbf{P}(t)} \min_{y \in \hat{\mathbf{V}}_{smpl}(t)} |x - y|_2^2, \quad (9)$$

4. FreeMotion Dataset

Previous LiDAR-related human motion datasets typically involve a single performer carrying out common actions with incomplete SMPL parameters, which have limitations in evaluating the generalization capability and robustness of HPS methods when being applied in daily-life complex scenarios. To facilitate the research of high-quality human motion capture in large-scale free environment, we provide FreeMotion, the first motion dataset with multi-view and multi-modal visual data with full-SMPL annotations, captured in diverse real-life scenarios with natural occlusions and noise. It contains 578,775 frames of data and annotations and contains $1 \sim 7$ performers in each scene.

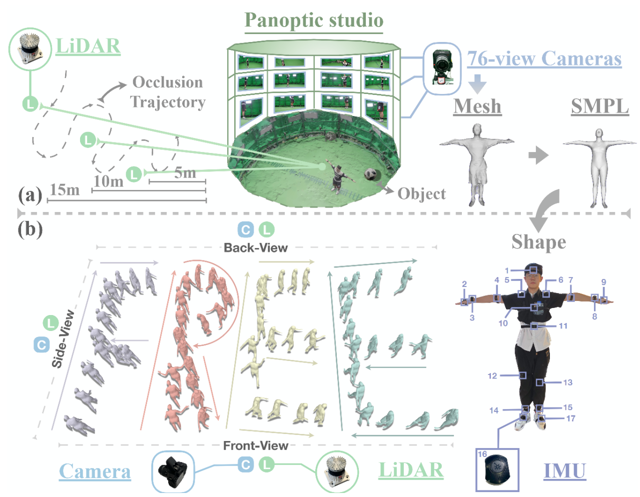

Figure 4. The capture systems of FreeMotion. In (a), we use a dense-camera capture system with LiDARs for accurate pose and shape capture. In (b), we set LiDARs and cameras at three views to capture human motions.

4.1. Data Acquisition

Considering that the indoor multi-camera panoptic studio can provide high-precision full SMPL parameter annotations and outdoor scenes are large-scale and suitable for real applications, we have two capture systems as shown in Fig. 4. For the first one, we set up a 76-Z-CAM system to obtain SMPL ground truth and three OUSTER-1 LiDARs at varied distances to get LiDAR data. Notably, we arrange other performers outside the studio to simulate occlusions in real-life scenarios. For the second one, we built three sets of LiDAR-camera capture devices, including a 128-beam OUSTER-1 LiDAR and a monocular Canon camera for each, in different locations to capture multi-view and multi-range visual data, and provide the global translation ground truth. The performer is equipped with a full set of Notiom equipment (17 IMUs) to obtain the pose ground truth. Particularly, the shape parameters of outdoor performers are captured in panoptic studio in advance. The capture frequencies for the LiDAR, Z-CAM, Canon camera, and IMU

Table 1. Comparison with public human motion datasets from four different aspects. "Capture distance" means the maximum distance between performer and capture device, which is approximately calculated with the data published. "Multi-person" indicates the capture scenes involve multiple persons. "HOI" denotes the human-object interaction scenarios.

Dataset	Statistics		Scenarios			Data				SMPL annotation		
	Frame	Capture distance(m)	Multi-person	In the wild	HOI	Point cloud	IMU	Image	Multi-view	Pose	Shape	Translation
AMASS [35]	16M	3.42	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓
HuMMan [6]	60M	3.00	✗	✗	✗	✓	✗	✓	✓	✓	✓	✓
SURREAL [48]	6M	N/A	✓	✗	✗	✗	✗	✓	✗	✓	✓	✓
AIST++ [31]	10M	4.23	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓
3DPW [52]	51k	N/A	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓
LiDARHuman26M [30]	184k	28.05	✗	✓	✗	✓	✓	✓	✗	✓	✗	✗
LIPD [42]	62k	30.04	✗	✓	✗	✓	✓	✓	✗	✓	✗	✗
MOVIN [20]	161k	N/A	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓
Sloper4D [11]	100k	N/A	✗	✓	✗	✓	✓	✓	✗	✓	✓	✓
CIMI4D [59]	180k	16.61	✗	✓	✗	✓	✓	✓	✗	✓	✓	✓
FreeMotion	578k	39.85	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

are set at 10Hz, 25Hz, 60Hz, and 60Hz, respectively. All the data are calibrated and synchronized.

4.2. Dataset Characteristics

The detailed comparison with existing public datasets is presented in Tab. 1. FreeMotion has several distinctive characteristics and we summarize three main highlights below.

Free Capture Scenes. Diverging from previous datasets focused on single-person HPS, FreeMotion is captured in real unconstrained environments, which involves diverse capture scales, multi-person activities, and human-object interaction scenarios. The large-scale human trajectories, occlusions, and noise all bring challenges for precise human global pose and shape estimation, thereby promoting the envelope of HPS technology for real-life applications.

Diverse Data Modalities and Views. FreeMotion offers multi-view and multi-modal capture data, including LiDAR point clouds, RGB images, and IMU measurements, providing rich resources for the exploration of single-modal, multi-modal, single-view, multi-view HPS solutions.

Complete Scene-level SMPL Annotations. Existing LiDAR-based motion datasets usually provide pose annotations using dense IMUs and lack annotations for human shape and global translation. FreeMotion remedies this by providing full SMPL parameters annotations, as shown in Fig. 4. FreeMotion involves 20 individuals with varying body types engaging in 40 types of actions. Details are in appendix. Accurate and complete annotations in rich scenarios can comprehensive evaluation for algorithms and benefit many downstream applications.

4.3. Data Extension

To enrich the dataset with various poses and shapes for pretraining, we follow LIP [42] to create synthetic point clouds from SURREAL [48], AIST++ [31], and portions of AMASS [35], including ACCAD and BMLMovi. It consists of 2,378k frames and 3,118 body shapes. Note that

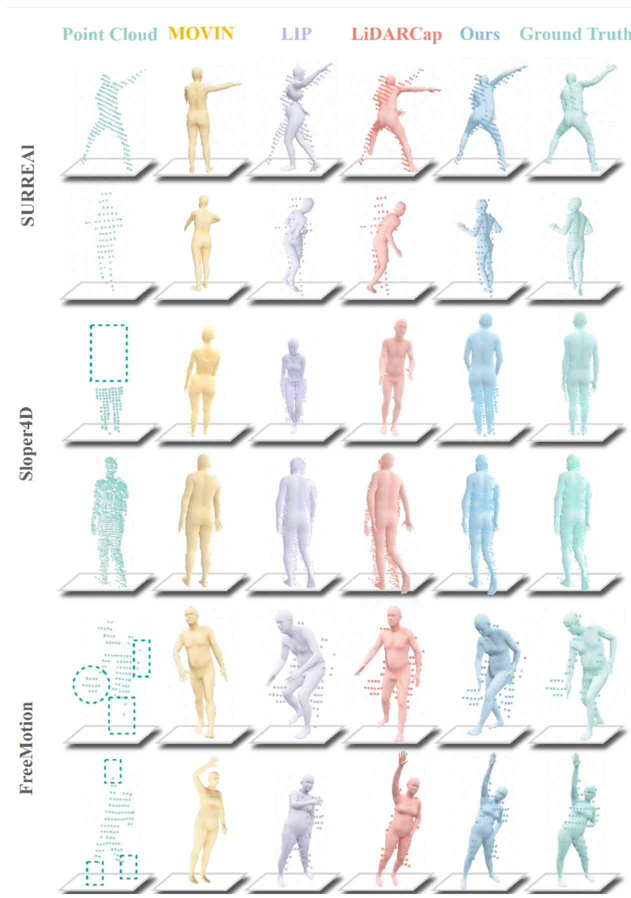

Figure 5. Qualitative comparisons. The point cloud matches the result better, representing more accurate estimation for pose, shape, and translation. Point cloud is far from results of MOVIN.

statistics in Tab. 1 do not include the synthetic data. *Detailed process is shown in appendix.*

5. Experiments

In this section, we compare our method with current SOTA methods on FreeMotion and various public datasets quali-

Table 2. Comparison with state-of-the-art methods on various datasets. Lower values represent better performance for all metrics.

	SURREAL [48]					Sloper4D [11]					FreeMotion				
	J/V Err(P)↓	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	SUCD↓	J/V Err(P)↓	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	SUCD↓	J/V Err(P)↓	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	SUCD↓
LiDARCap [30]	42.82/54.05	51.05/62.42	118.51/123.84	9.90	3.02	67.40/80.08	71.99/86.58	179.33/185.39	15.92	4.54	87.58/105.97	88.98/107.64	186.55/196.06	16.79	5.00
LIP [42]	31.72/42.40	32.71/43.02	45.22/53.18	12.17	0.95	60.11/74.90	62.03/77.26	94.81/106.34	19.95	2.27	81.13/98.65	81.99/99.58	129.88/141.77	18.76	4.24
MOVIN [20]	97.34/120.49	103.37/125.64	-	26.98	-	123.80/146.25	126.19/148.69	45559.25/45564.41	32.12	3311762.48	109.62/128.66	113.47/132.25	2853.87/2863.89	27.34	9252.86
LiveHPS	**23.99/30.81**	**24.75/31.81**	**34.45/40.14**	**9.49**	**0.67**	**46.22/56.72**	**48.28/59.02**	**77.73/85.83**	**12.77**	**1.67**	**68.88/83.20**	**69.43/83.90**	**119.27/128.61**	**15.79**	**2.85**

Table 3. The cross-dataset evaluation on various datasets. We use applicable metrics for each dataset according to its annotations.

	LiDARHuman26M [30]			LIPD [42]			CIMI4D [59]				SemanticKITTI [3]	HuCenLife [58]
	J/V Err(P)↓	Ang Err↓	SUCD↓	J/V Err(P)↓	Ang Err↓	SUCD↓	J/V Err(P)↓	J/V Err(PS)↓	Ang Err↓	SUCD↓	SUCD↓	SUCD↓
LiDARCap [30]	123.09/151.55	26.41	5.81	97.41/119.89	18.48	4.30	205.24/253.58	205.51/255.58	32.68	14.40	10.07	6.01
LIP [42]	103.48/124.18	24.14	3.93	83.38/102.25	18.73	2.81	162.28/205.25	166.38/211.10	33.03	8.47	9.93	5.06
MOVIN [20]	104.89/127.32	32.56	188906.16	101.78/121.67	28.82	66400.65	178.48/214.16	182.29/218.07	42.41	39681.84	1647852.73	58655.42
LiveHPS	**101.33/121.74**	**23.58**	**2.67**	**78.63/97.45**	**18.36**	**1.98**	**142.00/181.21**	**149.42/190.60**	**32.17**	**4.26**	**7.28**	**3.14**

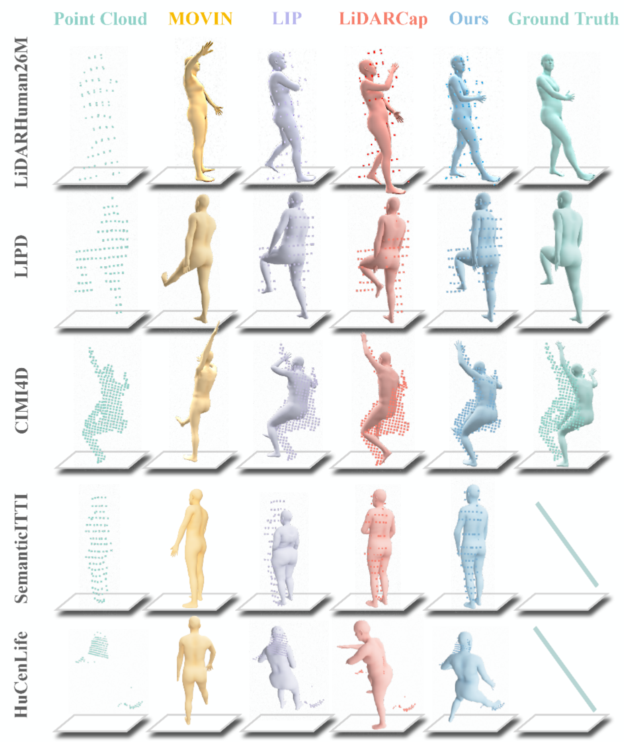

Figure 6. Qualitative comparisons in cross-dataset evaluation. SemanticKITTI and HucenLife do not provide SMPL annotations.

tatively and quantitatively, demonstrating our method's superiority and generalization capability. We also present detailed ablation studies for our network's modules to validate their effectiveness. Our evaluation metrics include 1) J/V Err(P/PS/PST) ↓: mean per joint/vertex position error in millimeters, where we generate joint/vertex from SMPL model by Pose/Pose-Shape/Pose-Shape-Translation parameters; 2) Ang Err ↓: mean per global joint rotation error in degrees to evaluate local pose; 3) SUCD ↓: scene-level unidirectional Chamfer distance in millimeters.

5.1. Implementation Details

We build our network on PyTorch 1.8.1 and CUDA 11.1, trained over 200 epochs with batch size of 32 and sequence length of 32, using an initial learning rate of 10^{-3}. The process was run on a server equipped with an Intel(R) Xeon(R) E5-2678 CPU and 8 NVIDIA RTX3090 GPUs. For training, we used clustered and manually annotated human point cloud sequences from raw data, while for testing, we employ sequential point clouds of human instances processed by a pre-trained segmentation model [53]. As for the dataset splitting, we take training set of FreeMotion, Sloper4D, and synthetic dataset including training set of SURREAL, AIST++, ACCAD, and BMLMovi for training.

5.2. Comparison

We evaluate LiveHPS against other state-of-the-art (SOTA) LiDAR-related methods [20, 30, 42] on FreeMotion and several public datasets [3, 11, 30, 42, 48, 58, 59] to demonstrate its superiority in capturing human global poses and shapes in large-scale free environment, even with severe occlusions and noise. Our LiveHPS achieves SOTA performance as shown in Tab. 2. The J/V Err(P) and Ang Err only relate to the pose parameter estimation, we surpass LiDARCap [30], LIP [42] and MOVIN [20] by an obvious margin. For fair comparison, we only use the LiDAR branch of LIP. As the pioneer to fully estimate SMPL parameters for LiDAR-based HPS, we develop a shape regression head with the same architecture of their pose regression head for fair comparison with other methods [20, 30, 42], the translation prediction of LiDARCap is the average of point cloud. Visual comparisons in Fig. 5 further highlight our method's superiority in global pose and shape estimations, yielding results that closely mirror ground truth. Other methods struggle in situations with occlusions and noise, as exemplified in challenging scenes from Sloper4D [11] and FreeMotion in Fig. 5. MOVIN [20] estimate translation based on velocity regression, it is not applicable on synthetic data SURREAL without real trajectories. Our LiveHPS demonstrates robust performance against noise like carried objects, as demonstrated in FreeMotion's left case in Fig. 5.

Tab. 3 illustrates our cross-dataset evaluation to validate the generalization capability of LiveHPS by directly testing on other datasets. LiDARHuman26M [30] and LIPD [42] only offer pose parameters. CIMI4D [59] provides pose, shape, and translation, but the translation is not that precise

Table 4. Ablation studies for our network modules. We also evaluate the internal details of each module.

	Network Module		Consecutive Pose Optimizer		Multi-head SMPL(Pose and Shape) Solver		Skeleton-aware Translation Solver			Ours
	w/o VAD	w/o CPO	w/o Temporal	w/o Spatial	ST-GCN	GRU	Average	MOVIN	LIP	
J/V Err(PST)	129.19/140.42	127.44/140.87	121.93/135.56	120.20/129.51	124.37/135.38	120.63/130.83	177.66/184.57	1296.48/1310.95	165.04/172.36	**119.27/128.61**
Ang Err	16.95	25.20	27.58	18.09	19.40	18.34	-	-	-	**15.79**
SUCD	3.17	4.08	3.51	2.95	3.28	3.08	4.25	8569.68	3.07	**2.85**

Table 5. More results on different lengths of input sequence and different point numbers on humans on FreeMotion dataset.

Frames	1	4	8	16	32
J/V Err(PST)↓	142.88/155.58	130.73/141.10	126.23/135.60	123.08/132.14	**119.27/128.61**
Ang Err↓	19.22	17.31	16.53	16.05	**15.79**
SUCD↓	5.22	3.03	3.01	3.02	**2.85**
Points	$0 \sim 100$	$100 \sim 200$	$200 \sim 300$	$300 \sim 1000$	> 1000
J/V Err(PST)↓	156.01/168.42	106.03/114.00	106.81/113.98	103.96/110.37	81.01/87.70
Ang Err↓	16.78	16.34	15.17	13.64	12.84
SUCD↓	4.54	2.31	2.25	2.52	2.63

as shown in the third row of Fig. 6. SemanticKITTI [3] and HuCenLife [58] are large-scale datasets for 3D perception, not providing SMPL annotations. Thanks to our VAD module's ability to harmonize diverse human point cloud distributions and CPO module's ability to model geometric and dynamic human features, our method can achieve SOTA performance on these cross-domain datasets, even in challenging cases with extreme occlusions, as Fig. 6 shows.

5.3. Ablation Study

We first validate the effectiveness of each module in LiveHPS. Then, we evaluate inner designs of each module to verify the effectiveness of detailed structures.

Tab. 4 shows the performance of our method with different network modules, demonstrating the necessity of our vertex-guided adaptive distillation (VAD) and consecutive pose optimizer (CPO) modules. We also illustrate ablation details of attention-based temporal and spatial feature enhancement in CPO, showing that the combination of temporal and spatial feature interaction performs best. We also conduct experiments to validate our attention-based multi-head SMPL solver. Our pose and shape solver, using the same network as CPO, outperforms ST-GCN from LiDARCap [30] and GRU from LIP [42] by fully utilizing the global temporal context and local spatial relationship existing in consecutive body joints. For the translation solver, the average of point cloud can reflect the coarse translation but it is very unstable with the distribution of point cloud changes. Compared with global velocity estimation utilized in MOVIN [20], our skeleton-aware translation solver directly estimates translations without error accumulation. Moreover, unlike GRU-based pose-guided corrector in LIP [42] which overlooks relationship between the skeleton and point cloud, our approach performs better by considering the relationship and more spatial information.

5.4. Generalization Capability Test

We assess the generalization capability of LiveHPS across varying lengths of input point cloud sequences and across different point numbers on human body in each frame, as Tab. 5 shows. Our method performs better with increasing sequence length but maintains good accuracy even with short inputs. In addition, our method performs relatively robust even for the situation with 100 points on the human body, which means far distance (about 15 meters) to LiDAR or severe occlusion. Fig. 1 and Fig. 7 show our method is practical for in-the-wild scenarios, capturing human motion in large-scale scenes day and night with real-time performance up to 45 fps. This strongly demonstrates the feasibility and superiority of our method in real-life applications. *More application results are in appendix.*

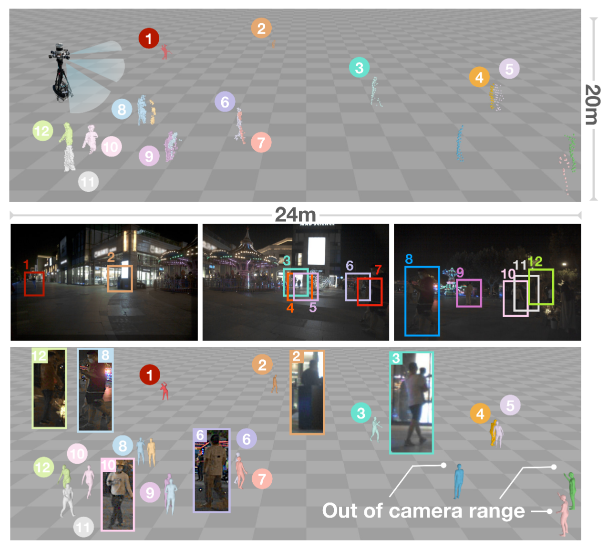

Figure 7. Performance of LiveHPS on real-time-captured scenes.

6. Conclusion

In this paper, we propose a novel single-LiDAR-based approach for predicting human pose, shape, and translations in large-scale free environment. To solve the occlusion and noise interference, we design a novel distillation mechanism and temporal-spatial feature interaction optimizer. Importantly, we propose a huge multi-person human motion dataset, which is significant for future in-the-wild HPS research. Extensive experiments on diverse datasets demonstrate the robustness and effectiveness of our method.

Limitations When human is static in the large-scale scene for a long time, our model can not fully utilize the dynamic information in consecutive frames and cause the misjudged orientation of human global orientations, opposite to the ground-truth pose.

References

[1] Sikander Amin, Mykhaylo Andriluka, Marcus Rohrbach, and Bernt Schiele. Multi-view pictorial structures for 3D human pose estimation. In *BMVC*, 2009. 2

[2] Andreas Baak, Meinard Müller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *ICCV*, 2011. 2

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 7, 8

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 2

[5] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 3D pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013. 2

[6] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pages 557–577. Springer, 2022. 3, 6

[7] Zhongang Cai, Liang Pan, Chen Wei, Wanqi Yin, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Pointhps: Cascaded 3d human pose and shape estimation from point clouds. *arXiv preprint arXiv:2308.14492*, 2023. 2

[8] Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, and Yuexin Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *CVPR*, pages 19608–19617, 2022. 2

[9] Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, and Yuexin Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. *arXiv preprint arXiv:2204.01026*, 2022. 1

[10] Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *CVPR*, pages 6792–6802, 2022. 3

[11] Yudi Dai, Yitai Lin, Xiping Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. *arXiv preprint arXiv:2303.09095*, 2023. 3, 6, 7

[12] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Mykhaylo Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, 2015. 2

[13] Kaiwen Guo, Jonathan Taylor, Sean Fanello, Andrea Tagliasacchi, Mingsong Dou, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Twinfusion: High framerate non-rigid fusion through fast correspondence tracking. In *3DV*, pages 596–605, 2018. 2

[14] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 38(2):14:1–14:17, 2019. 2

[15] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 1

[16] Yannan He, Anqi Pang, Xin Chen, Han Liang, Minye Wu, Yuexin Ma, and Lan Xu. Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In *CVPR*, pages 11400–11411, 2021. 1, 2

[17] 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 *3DV*, pages 421–430, 2017. 2

[18] 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 (TOG)*, 37(6):1–15, 2018. 2

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2013. 3

[20] Deok-Kyeong Jang, Dongseok Yang, Deok-Yun Jang, Byeoli Choi, Taeil Jin, and Sung-Hee Lee. Movin: Real-time motion capture using a single lidar. *arXiv preprint arXiv:2309.09314*, 2023. 2, 3, 6, 7, 8

[21] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 2

[22] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[23] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 2

[24] 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. *IRAL*, 4(2):1940–1947, 2019. 3

[25] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2

[26] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. Pare: Part attention regressor for

3d human body estimation. In *ICCV*, pages 11127–11137, 2021. 1, 2

[27] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 2

[28] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, pages 11605–11614, 2021. 1

[29] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, pages 6050–6059, 2017. 2

[30] Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. *arXiv preprint arXiv:2203.14698*, 2022. 2, 3, 5, 6, 7, 8

[31] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. 3, 6

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 1, 5

[33] Yuhang Lu, Qi Jiang, Runnan Chen, Yuenan Hou, Xinge Zhu, and Yuexin Ma. See more and know more: Zero-shot point cloud segmentation via multi-modal visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21674–21684, 2023. 1

[34] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[35] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 3, 6

[36] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516. IEEE, 2017. 3

[37] Noitom. Noitom Motion Capture Systems. https://www.noitom.com/, 2015. 1

[38] OptiTrack. OptiTrack Motion Capture Systems. https://www.optitrack.com/, 2009. 2

[39] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *CVPR*, 2017. 2

[40] Xidong Peng, Xinge Zhu, and Yuexin Ma. Cl3d: Unsupervised domain adaptation for cross-lidar 3d detection. *AAAI*, 2023. 2

[41] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, pages 11488–11499, 2021. 1

[42] Yiming Ren, Chengfeng Zhao, Yannan He, Peishan Cong, Han Liang, Jingyi Yu, Lan Xu, and Yuexin Ma. Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. *TVCG*, 2023. 2, 3, 5, 6, 7, 8

[43] Helge Rhodin, Nadia Robertini, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *ICCV*, 2015. 2

[44] Nadia Robertini, Dan Casas, Helge Rhodin, Hans-Peter Seidel, and Christian Theobalt. Model-based outdoor performance capture. In *3DV*, 2016. 2

[45] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 2

[46] Leonid Sigal, Alexandru O. Bălan, and Michael J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 2010. 3

[47] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2

[48] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 6, 7

[49] Vicon. Vicon Motion Capture Systems. https://www.vicon.com/, 2010. 2

[50] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. Practical motion capture in everyday surroundings. *TOG*, 26(3):35–es, 2007. 2

[51] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, pages 349–360. Wiley Online Library, 2017. 2

[52] 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 *ECCV*, pages 601–617, 2018. 3, 6

[53] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 7

[54] Xiaolin Wei, Peizhao Zhang, and Jinxiang Chai. Accurate realtime full-body motion capture using a single depth camera. *SIGGRAPH Asia*, 31(6):188:1–12, 2012. 2

[55] XSENS. Xsens Technologies B.V. https://www.xsens.com/, 2011. 1, 2

[56] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *CVPR*, 2020. 2

[57] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 37 (2):27:1–27:15, 2018. 2

[58] Yiteng Xu, Peishan Cong, Yichen Yao, Runnan Chen, Yuenan Hou, Xinge Zhu, Xuming He, Jingyi Yu, and Yuexin Ma. Human-centric scene understanding for 3d large-scale scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20349–20359, 2023. 7, 8

[59] Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. *arXiv preprint arXiv:2303.17948*, 2023. 3, 6, 7

[60] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 1, 2

[61] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *CVPR*, 2022. 1, 2

[62] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. *CVPR*, 2021. 2

[63] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. *TPAMI*, 2019. 2

[64] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3d human pose and shape. *arXiv preprint arXiv:2008.06910*, 2020. 2

[65] Xinge Zhu, Yuexin Ma, Tai Wang, Yan Xu, Jianping Shi, and Dahua Lin. Ssn: Shape signature networks for multi-class object detection from point clouds. In *ECCV*, pages 581–597. Springer, 2020. 1, 2

[66] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Wei Li, Yuexin Ma, Hongsheng Li, Ruigang Yang, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. *TPAMI*, 2021. 1, 2