LiveHPS++: Robust and Coherent Motion Capture in Dynamic Free Environment

Yiming Ren[1], Xiao Han[1], Yichen Yao[1], Xiaoxiao Long[2], Yujing Sun[2⋆], and Yuexin Ma[1⋆]

[1]ShanghaiTech University, [2]The University of Hong Kong
{renym2022,mayuexin}@shanghaitech.edu.cn

Fig. 1: Visualization of the motion capture performance of LiveHPS++ in a real-time captured scenario with complex human-object interaction. The left exhibits images for reference, the middle shows the noised point clouds (top) and our corresponding mesh model results (bottom). We zoom in some cases on the right for clearer demonstration, where point clouds are drawn in white.

Abstract. LiDAR-based human motion capture has garnered significant interest in recent years for its practicability in large-scale and unconstrained environments. However, most methods rely on cleanly segmented human point clouds as input, the accuracy and smoothness of their motion results are compromised when faced with noisy data, rendering them unsuitable for practical applications. To address these limitations and enhance the robustness and precision of motion capture with noise interference, we introduce LiveHPS++, an innovative and effective solution based on a single LiDAR system. Benefiting from three meticulously designed modules, our method can learn dynamic and kinematic features from human movements, and further enable the precise capture of coherent human motions in open settings, making it highly applicable to real-world scenarios. Through extensive experiments, LiveHPS++ has proven to significantly surpass existing state-of-the-art methods across various datasets, establishing a new

⋆ Corresponding author. This work was supported by NSFC (No.62206173), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

benchmark in the field. https://4dvlab.github.io/project_page/LiveHPS2.html

1 Introduction

Capturing accurate and natural human motions in large-scale dynamic environments is pivotal for the modeling and analysis of human behaviors, as well as for enhancing the understanding of 3D scenes. This foundational work significantly benefits many downstream applications, ranging from digital filmmaking and delivering immersive experiences in AR/VR gaming to training robots to emulate human-like behaviors and effectively collaborate with humans. Previous motion capture methods are usually based on optical devices [1,5,12,31,35,36,40] or wearable devices [29,48,54,55]. However, the former is sensitive to light conditions and not suitable for outdoor scenarios, and the latter requires the actor to wear a body-mounted IMU suit, not applicable for daily-life usage and fails to capture human shapes.

LiDAR has emerged as a foundational sensor in the realms of robotics and autonomous driving [8, 53, 60, 61], due to its exceptional long-range depth-sensing capabilities. Notably, LiDAR point clouds can offer precise 3D geometry and location information of the human body without limitation of light conditions or wearable devices. This makes LiDAR very promising for tracking how people move and act in free environment. Recently. some advancements [6, 20, 27] have already underscored the efficacy of single-LiDAR systems for capturing human motion. However, these techniques are primarily effective with clean human point clouds within controlled experimental settings, and they often fall short when confronted with the intricacies of real-world application scenarios with noise and occlusions.

To address these problems, LiveHPS [33] has collected a vast dataset of human motion captured in real-world settings, featuring natural interactions with other people, and has also put forth an effective strategy to solve the varying point distributions stemming from occlusions and noise. Nevertheless, it treats features from real human points and noise points equally and ignores the coherence of global poses and translations, causing the accuracy and consistency of its predicted motions to face limitations in complex scenarios, where the noise does not just come from the sensor accuracy but also arises from the horrible segmentation results by upstream perception algorithms. As illustrated in Fig.1, distinguishing clean human point clouds becomes particularly challenging when individuals are in close proximity or interacting with objects. This usually results in disastrous inputs with much noise for motion capture algorithms, leading to inaccurate and jerky motion outcomes. Enhancing the robustness and precision of LiDAR-based motion capture methods in any complicated situation becomes very crucial for real-world deployment and applications.

In this paper, we introduce **LiveHPS++**, an innovative and effective approach for capturing precise and coherent human motions across vast, unregulated environments using a single LiDAR system. Our method consists of three specially designed modules to tackle above challenges. Beginning with sequential point cloud inputs, the first module, the **Trajectory-guided Body Tracker**, subtly captures the dynamic characteristics of human movement through their trajectories, ensuring consistency across adjacent

poses. Following this, the second module, the **Noise-insensitive Velocity Predictor**, employs a cross-attention mechanism to make each human joint engage with the most relevant point features, minimizing the impact of extraneous noise. This module forecasts the velocity of each joint, explicitly modeling the kinematic details of the human motion, which is very valuable for refining the global pose and translation in the third module, the **Kinematic-aware Pose Optimizer**. Here, a sophisticated system of candidate generation and feature interaction is implemented to achieve more precise pose adjustments. In particular, we propose a synthesized motion dataset, **NoiseMotion**, built on SURREAL [43] and ShapeNet [7] to enlarge the challenging noised data in complex scenarios by simulating human interactions with various objects. Benefiting from both the new dataset and our effective algorithm, we significantly enhance the robustness and accuracy of motion capture in complex settings, where noise interference is prevalent, thereby facilitating the practical deployment of this technology.

Extensive experiments are conducted on multiple LiDAR-based motion datasets, including NoiseMotion, FreeMotion [33], FreeMotion-OBJ [33], and Sloper4D [10]. Compared with the state-of-the-art method, LiveHPS++ has improved the performance by a large margin, e.g. 6.28% and 69.29% for the global vertex error and the jitter on FreeMotion-OBJ, the most challenging real human motion dataset, and 23.05% and 13.54% for the global vertex error and the jitter on NoiseMotion, the most challenging synthetic human motion dataset. Moreover, detailed ablation studies and discussions are also provided to verify the effectiveness of detailed designs in our network. Contributions of this work can be summarized as follows.

- We propose a robust human motion capture method, LiveHPS++, which can eliminate the effect of severe noises and produce precise and natural human motions in dynamic free environment, which is very practical for real-world applications.
- We design three effective modules in our method, which can implicitly and explicitly model dynamic and kinematic features of human motions to facilitate the coherence and accuracy of motion capture results.
- LiveHPS++ achieves state-of-the-art performance on various datasets and significantly outperforms existing methods.

2 Related Work

2.1 Optical-based Motion Capture

Early systems of optical-based motion capture, characterized by their reliance on camera-tracked markers to reconstruct 3D meshes [30, 44, 45], laid the groundwork for high-quality motion capture, becoming a staple in professional settings. The field has seen substantial advancements with the introduction of markerless mocap technologies [4, 11, 17, 21, 28, 38, 39, 41, 42, 49, 50, 58]. These innovations offer a less intrusive and often more cost-effective solution, leveraging multi-view algorithms to maintain robustness even in uncontrolled environments [1, 5, 12, 31, 35, 36, 40]. However, the complexity of synchronizing and calibrating multi-camera setups remains a challenge. In response, monocular mocap methods have been developed, employing a range of techniques from

optimization and regression [3, 18, 22–26, 59] to template-based and probabilistic approaches [14–16, 51, 52]. To address depth ambiguity, some researchers have turned to depth cameras [2, 13, 37, 47, 57], which, despite their benefits, are limited by their sensing range and perform poorly in outdoor settings. Nonetheless, optical-based motion capture methods remain constrained by lighting conditions, limiting their applicability in outdoor scenes.

2.2 Inertial-based Motion Capture

Inertial motion capture systems offer a distinct advantage over traditional optical systems by being impervious to occlusions and unrestricted by lighting conditions or recording environment volume. However, the commercial solutions typically require performers to wear form-fitting suits equipped with a large number of IMUs, leading to setups that are intrusive and cumbersome for the wearer. Some recent methods [19, 34, 46, 54, 55] utilize sparse IMUs to produce promising results. Despite it improves the portability of actors during motion capture, it still suffers from drift errors over time and is unable to accurately perceive other physical information, such as human shape and translations.

2.3 LiDAR-based Motion Capture

LiDAR technology, known for its precise long-range depth-sensing capabilities, has become increasingly pivotal in fields such as robotics and autonomous driving [9, 32, 56, 60, 61]. Its ability to deliver accurate depth information across expansive environments, irrespective of lighting conditions, marks it as an invaluable tool for robust 3D Human Pose and Shape (HPS) estimation. PointHPS [6] showcases the potential of using cascaded network architectures for pose and shape estimation directly from point clouds, but the network architecture relies on dense point cloud inputs and is not suitable for sparse point cloud data captured in outdoor large-scale scenes. LiDARCap [27] introduces a graph-based convolutional network approach tailored for interpreting daily human poses within the vast and variable scales of LiDAR-captured scenes. MOVIN [20] explores generative methods for human pose and global translation estimation. However, these methods focus exclusively on human point clouds in noise-free environments. Recently, LiveHPS [33] propose a scene-level human pose and shape estimation by fully utilizing the temporal and spatial information to solve the occlusion and noise disturbance. However, the network tends to take features of all points, including both true human points and noise points, as valid information. The severe noise in data will greatly affect the accuracy of results. Moreover, it only contains the interaction between joints, ignoring the global kinematic information, leading the incoherence in global motion capture outcomes. Advancing into this research domain, we aim to fully exploit the dynamic and kinematic information available in human movements to capture more coherent and accurate global human motions in noise environments.

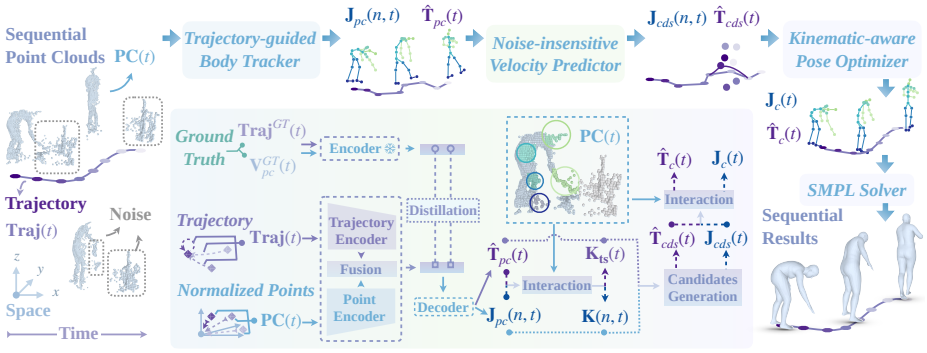

Fig. 2: The pipeline of LiveHPS++. It consists of three primary modules, including a trajectory-guided body tracker to predict the human joint and translation, a noise-insensitive velocity predictor to regress the velocity, and the kinematic-aware pose optimizer to enhance the accuracy and coherence of results. Finally, we use SMPL solver to regress the parameters of human poses and shape. Detailed network structure of three modules is also shown under the upper pipeline.

3 Methodology

We introduce LiveHPS++, a novel single-LiDAR-based methodology to estimate the robust and coherent human motions in dynamic free environment. The overview of the pipeline is shown in Fig. 2. It takes the sequential noise point clouds as input and aims to acquire the sequential SMPL parameters, including human poses, shapes and translations. The pipeline is structured in three critical components, including trajectory-guided body tracker(Sec 3.2), noise-insensitive velocity predictor(Sec 3.3), and kinematic-aware pose optimizer(Sec 3.4). Firstly, we employ trajectory-guided body tracker to predict the human joints and translations with the assistance of global dynamic information. Secondly, we propose the noise-insensitive velocity predictor to fully utilize the potential association between the human joints and original point cloud for regressing the velocity of each global joint, aiming to optimize the results and concurrently mitigate noise impacts. Then, we design the kinematic-aware pose optimizer to enhance the coherence and accuracy of human motion by predicted velocity. Finally, we use SMPL solver to predict the human poses and shapes from the coherent human joints.

3.1 Preliminaries

LiveHPS++ takes sequential single human point clouds as input interspersed with noise from surrounding objects. We resample each input point cloud to a fixed $\mathbf{N}_{input} = 256$ by farthest point sample algorithm, then we subtract point clouds with each average location and record the location $\mathbf{Loc}(t) \in R^3$. We define $\boldsymbol{\theta}^{GT}(t) \in \mathbf{R}^{6N_J}$, $\boldsymbol{\beta}^{GT}(t) \in \mathbf{R}^{10}$, and $\mathbf{T}^{GT}(t) \in \mathbf{R}^3$ as the ground truth SMPL parameters, $\mathbf{N}_J = 24$ and $\mathbf{N}_V = 6890$ represents the number of human joint and mesh vertex. We define $\mathbf{PC}(t)$ and $\mathbf{Loc}(t)$ as the normalized point cloud and the mean average positions of raw point cloud, we also follow LIP [34] to simplify the translation prediction as the offset prediction $\hat{\mathbf{T}}(t)$, and define $\hat{\mathbf{T}}^{GT}(t)$ as the ground truth, the equation is formu-

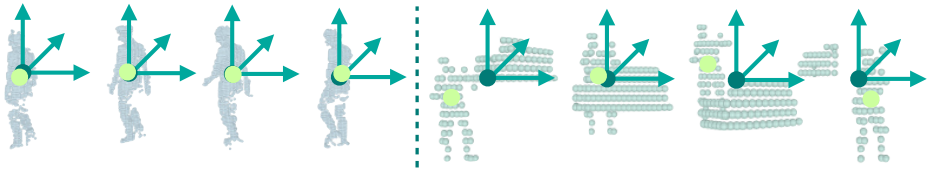

Fig. 3: Normalized point cloud. The light green point represents the human root positions, while the dark green point represents the origin of coordinate axis after normalization. The sequential point cloud on the left without noise can be normalized to obtain a relatively stable data distribution, while the data on the right exhibits a more jittery data distribution after normalization due to noise interference.

lated below:

$$\hat{\mathbf{T}}^{GT}(t) = \mathbf{Loc}(t) - \mathbf{T}^{GT}(t). \tag{1}$$

We also define the $\mathbf{K}^{GT}(n)$ and $\mathbf{K}_{ts}^{GT}$ as the velocity supervision, the equation is formulated below:

$$\begin{aligned} \mathbf{K}^{GT}(n,t) =& \mathbf{J}^{GT}(n,t+1) - \mathbf{J}^{GT}(n,t), \\ \mathbf{K}_{ts}^{GT}(t) =& \mathbf{T}^{GT}(t+1) - \mathbf{T}^{GT}(t). \end{aligned} \tag{2}$$

3.2 Trajectory-guided Body Tracker

The general strategy of normalization for input data is subtracting the average location, which aims to enhance the generalization capabilities of network and maintain stability in the input data. Notably, the vertex-guided adaptive distillation mechanism, as proposed by LiveHPS [33], relies heavily on this normalized input to achieve point representations that closely align with the ground truth vertices, thereby facilitating more accurate and consistent motion capture. However, this conventional normalization strategy encounters significant challenges when dealing with dynamic noise point cloud data. Noise introduced by objects or backgrounds in the scene can lead to substantial fluctuations in the point cloud distribution between adjacent frames, as shown in Fig. 3. Such fluctuations can disrupt the spatial continuity of the temporal trajectory information, leading to instability in the input data. To restore stability in the normalized data amidst noise disruptions, we introduce a dedicated encoder designed to capture trajectory embedding to implicitly model the dynamic characteristics of human movement. We also modify the mechanism as **vertex-trajectory-guided adaptive distillation** with extra ground truth trajectory information, aiming to fully preserve the trajectory information, thereby enabling to capture of more precise and coherent human motion. Additionally, the transformation of point cloud features into vertex features in a high-dimensional space potentially predicts the displacement between the point cloud's average position and the true central point. Consequently, we incorporate a decoder branch specifically for predicting translations, further refining the accuracy of our motion capture process.

The distillation mechanism consists of two networks with the same architecture and different input data. We follow LiveHPS to generate the sampling of vertices $\mathbf{V}_{pc}^{GT}(t)$ with consistent distribution with input point cloud and generate the trajectory $\mathbf{Traj}^{GT}(t)$ by Equation 5 with $\mathbf{T}^{GT}(t)$. The guidance network takes $\mathbf{V}_{pc}^{GT}(t)$ and $\mathbf{Traj}^{GT}(t)$ as input, we use an MLP encoder to extract the trajectory feature and the PointNet-GRU

structure to extract the global point feature. Then we fuse above two features by an MLP layer to get fusion feature $\mathbf{F}_{gt}(t) \in \mathbb{R}^{1024}$ and predict the translations $\hat{\mathbf{T}}_{gt}(t)$ and human joints $\mathbf{J}_{gt}(t)$ by an MLP decoder. We train the guidance network by the mean squared error loss for supervision and freeze the parameters.

$$\mathcal{L}_{mse}(\mathbf{J}_{gt}) = \sum_{t} \parallel \mathbf{J}_{gt}(t) - \mathbf{J}^{GT}(t) \parallel_2^2, \tag{3}$$

$$\mathcal{L}_{mse}(\mathbf{T}_{gt}) = \sum_{t} \parallel \hat{\mathbf{T}}_{gt}(t) - \hat{\mathbf{T}}^{GT}(t) \parallel_2^2 . \tag{4}$$

We record the average location $\mathbf{Loc}(t)$ of input data and calculate the trajectory $\mathbf{Traj}(t)$ relative to the first frame input data.

$$\mathbf{Traj}(t) = \mathbf{Loc}(\mathbf{t}) - \mathbf{Loc}(\mathbf{1}). \tag{5}$$

The learning network takes the input point cloud $\mathbf{PC}(t)$ and trajectory $\mathbf{Traj}(t)$ as input and follows above steps to get the fusion feature $\mathbf{F}_{pc}(t)$ and predict the translations $\hat{\mathbf{T}}_{pc}(t)$ and human joints $\mathbf{J}_{pc}(t)$. The loss function of the trajectory-guided body tracker(TBT) $\mathcal{L}_{TBT}$ is formulated as below:

$$\mathcal{L}_{distillation} = \sum_{t} \mathbf{F}_{gt}(t) \log(\frac{\mathbf{F}_{gt}(t)}{\mathbf{F}_{pc}(t)}), \tag{6}$$

$$\mathcal{L}_{TBT} = \lambda_1 \mathcal{L}_{distillation} + \lambda_2 \mathcal{L}_{mse}(\mathbf{J}_{pc}) + \lambda_3 \mathcal{L}_{mse}(\hat{\mathbf{T}}_{pc}), \tag{7}$$

where $\lambda_1 = 10^3$, $\lambda_2 = 1$ and $\lambda_3 = 1$ are hyper-parameters. During inference, the guidance network is not required.

3.3 Noise-insensitive Velocity Predictor

The trajectory-guided body tracker regresses human joint positions, which leverages skeletal geometric information to enhance motion capture accuracy. This parent-children joint structure allows for the correction of mispredicted joints when partial point cloud data is missing. However, the dependency can lead to error accumulation if the parent joint's prediction is skewed by noisy point cloud data, cause the algorithm noise. To address the challenge, we aim to enhance the motion feature and learn the kinematic expressions to eliminate the impact of noise. As shown in Fig. 2, we design the noise-insensitive velocity predictor to predict the velocity of each human joint which can reflect the kinematic information of human motions and further refine the global pose and translation.

Specifically, the module takes the human joints/translation and input point cloud as the input, utilizes the cross-attention mechanism to make each joint search for truly valuable point features from the original point cloud for feature enhancement, and predicts the velocity $\mathbf{K}(n)/\mathbf{K}_{ts} \in \mathbf{R}^L$ ($L = 32$ represents the temporal window size). The loss function is formulated as below:

$$\begin{aligned} \mathcal{L}_{mse}(\mathbf{K}(n)) &= \sum_{n} \parallel \mathbf{K}(n) - \mathbf{K}^{GT}(n) \parallel_2^2, \\ \mathcal{L}_{mse}(\mathbf{K_{ts}}) &= \parallel \mathbf{K}_{ts} - \mathbf{K}_{ts}^{GT} \parallel_2^2 . \end{aligned} \tag{8}$$

By supervision, the network can learn to distinguish features between real human points and noise points and then eliminate the noise effect.

3.4 Kinematic-aware Pose Optimizer

Leveraging the velocity values derived previously, we develop a kinematic-aware optimizer to further correct motion outcomes. The predicted velocity provides the temporal connection of each joint between $t+1$ frame and t frame, while the predicted joints $\mathbf{J}_{pc}(n,t)$ provide joint-wise spatial position of each frame, thus we can generate candidate joints $\mathbf{J}_{cds}(n,t,t)$ by:

$$\begin{aligned} \mathbf{J}_{cds}(n_i,t_i,t_j) &= \mathbf{J}_{pc}(n_i,t_i) + \Delta t \sum_{t=t_i}^{t_j} \mathbf{K}(n_i,t), \\ \hat{\mathbf{T}}_{cds}(t_i,t_j) &= \mathbf{Loc}(t) - (\mathbf{T}_{pc}(n_i,t_i) + \Delta t \sum_{t=t_i}^{t_j} \mathbf{K}_{ts}(t)). \end{aligned} \tag{9}$$

$\mathbf{J}_{cds}(n_i,t_i,t_j)$ represents the t_i candidate joints for n_i joint in t_j frame, thereby we can generate $(L-1)$ candidate joints. Considering that long-term kinematic optimization causes inaccurate results at the extreme points of one motion sequence and causes cumulative errors over time, meanwhile, short-term kinematic optimization with short-change information can only refine minority mutation cases and can not keep the coherence of the whole sequence. We design the module with cross-attention architecture to build the connection between the candidate joints with the original input data, facilitating the extraction of more pronounced features and further benefiting more accurate joint correction. Finally, we get the coherent and accurate global human joints $\mathbf{J}_c(t)$ and translations $\mathbf{T}_c(t)$, the loss function of the kinematic-aware pose optimizer(KPO) $\mathcal{L}_{KPO}$ is formulated as below:

$$\mathcal{L}_{KPO} = \lambda_4 \mathcal{L}_{mse}(\mathbf{J}_c) + \lambda_5 \mathcal{L}_{mse}(\hat{\mathbf{T}}_c), \tag{10}$$

where $\lambda_4 = 1$ and $\lambda_5 = 1$ are hyper-parameters.

3.5 SMPL Solver

In the last stage, we follow the LiveHPS [33] to use the attention-based SMPL solver to predict the human poses $\theta(t)$ and shape β. Finally, we use SMPL model to generate the human joints and mesh vertex as below:

$$\hat{\mathbf{J}}_{smpl}(t), \hat{\mathbf{V}}_{smpl}(t) = \mathrm{SMPL}(\theta(t), \beta, \hat{\mathbf{T}}_c(t)). \tag{11}$$

The loss function of the SMPL solver $\mathcal{L}_{smpl}$ is formulated as:

$$\begin{aligned} \mathcal{L}_{smpl} =& \lambda_6 \mathcal{L}_{mse}(\mathbf{J}_{smpl}) + \lambda_7 \mathcal{L}_{mse}(\mathbf{V}_{smpl}) \\ &+ \lambda_8 \mathcal{L}_{mse}(\theta) + \lambda_9 \mathcal{L}_{mse}(\beta), \end{aligned} \tag{12}$$

where $\lambda_6 = \frac{100}{N_j}$, $\lambda_7 = \frac{100}{N_v}$, $\lambda_8 = 1/5$ and $\lambda_9 = 1$ are hyper-parameters. It is worth noting that due to the large number of noise points in the input data, the SUCD loss proposed by LiveHPS is not suitable.

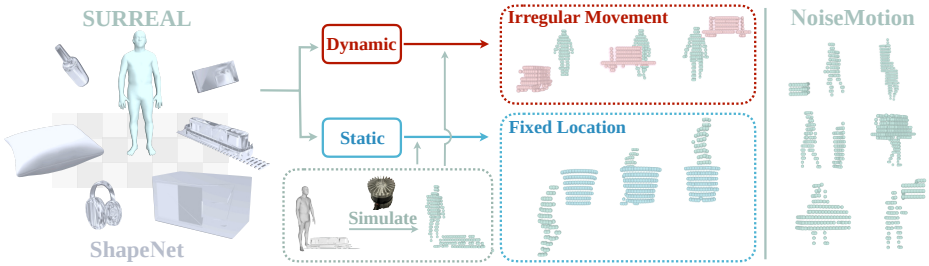

Fig. 4: The NoiseMotion dataset simulation pipeline, integrating dynamic human motion and static object noise to simulate real-world human-object interactions.

4 Dataset

Previous LiDAR-based methods use extensive synthetic data in training to enhance the generalization of network, however, existing synthetic data primarily simulate arbitrary sensing noise and exhibit random human point cloud translations across sequences, failing to accurately reflect the complexities encountered in real-world scenarios. Acknowledging the limitations of existing synthetic datasets, we propose the NoiseMotion, which leverages human motion data from SURREAL [43] and 3D object models from ShapeNet [7] to meticulously simulate complex noise patterns resulting from human-object interactions. The pipeline of data augmentation is shown in Fig. 4, which categorizes noisy objects as either dynamic or static. Dynamic objects, which can appear unexpectedly and significantly alter the point cloud distribution, contrast with static objects that introduce noise through human-object interaction or proximity, a prevalent noise type in everyday settings. To further diversify the noise types and their distribution, we employ data augmentation techniques like random rotation and scaling. This approach is critical for accurately reflecting real-world complexities, significantly boosting the network's generalization capability. Compared to the real human motion dataset FreeMotion-OBJ [33], NoiseMotion offers a vastly richer collection, including 51,300 unique 3D object models from ShapeNet and 1,021,802 human motions from SURREAL. In contrast, FreeMotion-OBJ provides fewer than ten types of dynamic objects. This stark difference underscores NoiseMotion's importance and necessity for advancing LiDAR-based applications.

5 Experiment

In this section, we present comprehensive experiments to validate the effectiveness, robustness, and coherence of our method, LiveHPS++, against current state-of-the-art (SOTA) methods, including LiDARCap [27], LIP [34], and LiveHPS [33]. Additionally, we present detailed ablation studies to assess the contribution of our network architecture's components. Following LiveHPS, our evaluation metrics include J/V Err(PS/PST)(mm) and Ang Err($degree$). Notably, scene-level unidirectional Chamfer distance in millimeters(SUCD) is not suitable for noisy input data, so we don't use SUCD for metrics. To provide deeper insight into the effect of our method on the accuracy and coherence of human motion trajectories, we introduce two additional metrics:

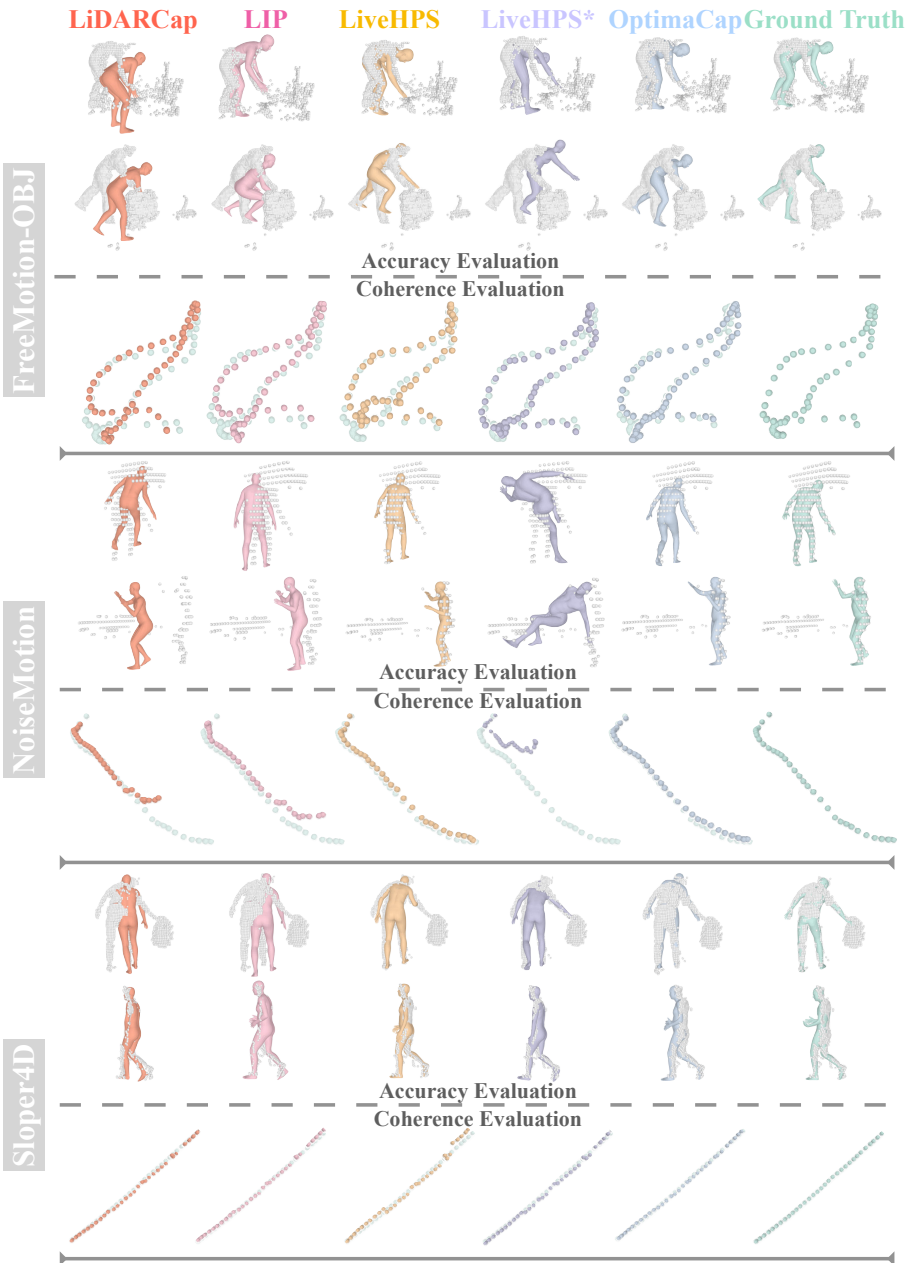

Fig. 5: Qualitative comparisons. The point cloud matches the result better, representing more accurate estimation for pose, shape, and translation. Each point in the visualization of coherence evaluation represents the frame-wise global human translations in the bird's-eye view.

1) Acceleration Error(Accel Err)(m/s^2)↓: which quantifies the mean acceleration error across global human joints, calculated against ground truth data to gauge the trajectory accuracy of human motion; 2) Jitter($10^2 m/s^3$)↓: which evaluates the average jerk

Table 1: Comparison with state-of-the-art methods on various datasets. Lower values represent better performance for all metrics. FreeMotion-OBJ means the human-object interaction part of FreeMotion. Notably, LiveHPS* is trained on the real dataset and previous clean synthetic data, which contains the same human motion as the NoiseMotion.

	NoiseMotion [43] [7]					FreeMotion-OBJ [33]				
	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	Jitter↓	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	Jitter↓
LiDARCap [27]	52.63/64.65	400.66/402.58	10.87	42.48	765.89	84.11/100.61	181.82/189.32	16.61	7.21	62.47
LIP [34]	62.41/77.97	192.79/198.66	14.07	25.31	451.74	87.50/108.28	158.38/170.90	20.16	7.09	60.19
LiveHPS [33]	48.37/60.42	74.70/83.84	12.19	5.78	68.65	70.73/88.43	146.78/158.00	17.81	8.82	117.79
LiveHPS* [33]	370.29/432.93	561.49/611.40	27.32	49.74	884.24	83.33/101.70	133.82/146.12	16.84	8.38	100.82
Ours	**34.00/42.75**	**58.53/64.51**	**10.63**	**3.48**	**59.35**	**58.11/72.55**	**128.60/136.94**	**15.85**	**7.01**	**30.96**
	FreeMotion [33]					Sloper4D [10]				
	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	Jitter↓	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	Jitter↓
LiDARCap [27]	86.28/104.17	180.36/188.58	15.51	6.28	70.57	71.64/84.23	138.71/147.79	13.72	6.16	88.50
LIP [34]	85.49/104.05	141.36/153.25	19.73	6.16	68.11	74.38/91.89	134.69/146.90	20.53	6.59	96.16
LiveHPS	74.71/90.79	130.41/141.08	16.96	7.27	85.38	53.37/63.15	88.35/95.85	13.08	5.88	73.56
LiveHPS*	69.38/83.86	119.22/128.55	15.80	6.99	86.07	48.28/59.02	77.73/85.83	12.77	5.64	97.41
Ours	**61.91/75.27**	**112.13/120.39**	**15.40**	**5.42**	**33.16**	**42.70/50.62**	**76.98/81.67**	**11.92**	**4.34**	**59.97**

across global human joints, assessing coherence of motion trajectories independently of ground truth data and offering a measure of the fluidity of captured movements.

5.1 Implemantation Details

Our network structure is implemented by PyTorch version 1.10.0 and CUDA 11.4, trained over 200 epochs with batch size of 64, sequence length of 32, and learning rate of 10^{-3}, on an Intel(R) Xeon(R) Gold 5318Y CPU and 4 NVIDIA A40 GPUs. Other training configuration aligns with the settings established by LiveHPS. As for the dataset, we use FreeMotion [33], Sloper4D [10], and our NoiseMotion. FreeMotion and Sloper4D are LiDAR-based human motion datasets, some of the data contains noise points from objects. Our NoiseMotion dataset is based on the SURREAL [43] and ShapeNet [7], we generate the synthetic data consisting of human motion dataset SURREAL and object dataset ShapeNet, which simulate the challenge human-object interaction case in dynamic free environment with severe noise. The dataset splitting is followed by LiveHPS. All methods are trained on the training set of NoiseMotion, FreeMotion, and Slopher4D.

5.2 Comparison

We evaluate LiveHPS++ on the testing sets of NoiseMotion, FreeMotion, and Sloper4D. To specifically assess our network's resilience to noise, we additionally show the evaluation results on the human-object interaction sequences from the FreeMotion testing set, referred to as FreeMotion-OBJ. This allows for a focused assessment of noise-handling capabilities. LiveHPS++ is benchmarked against leading LiDAR-based methods to underscore its state-of-the-art (SOTA) performance, as detailed in Tab. 1. The experimental results demonstrate our LiveHPS++'s exceptional performance across various metrics, with notable advancements in Acceleration Error (Accel Err) and Jitter especially in NoiseMotion, underscoring our method's adeptness at handling dynamic motion coherence in complex environments with severe noise. We can observe that,

after training on NoiseMotion, LiveHPS exhibits enhanced performance on NoiseMotion and FreeMotion-OBJ dataset characterized by high noise levels compared with LiveHPS*, which is trained on real data and previous clean synthetic dataset, which provides the same human motion as the NoiseMotion. This demonstrates the value of our synthetic data NoiseMotion. However, LiveHPS can not achieve stable generalizability and shows a relative decline in FreeMotion and Sloper4D when compared to LiveHPS*. Our LiveHPS++ can achieve SOTA performance in both cases with severe noise or not, highlighting the robustness and generalization capabilities of our method.

The qualitative comparisons in Fig. 5 further accentuate LiveHPS++'s proficiency in sustaining stability and delivering coherent outcomes under significant noise conditions. Through visualizations of global human motion by each method on the test datasets, LiveHPS++'s superior noise immunity is evident. For instance, in challenging scenarios such as those presented in FreeMotion-OBJ and NoiseMotion, our LiveHPS++ reliably differentiates between human-related and noise points, unlike LiveHPS*, which misinterprets noise as legitimate motion cues in noise environments, leading to inaccuracies. While other methods trained on NoiseMotion dataset show some capacity to disregard irrelevant noise, their performance is still noticeably impacted. Moreover, in scenarios with occlusion, such as those within the Sloper4D's second row where the hand is occluded, LiveHPS++ consistently outperforms competing methods in these scenarios. As Fig. 1 shows, thanks to our effective network design which can implicitly and explicitly model dynamic and kinematic features of human motions, our LiveHPS++ can capture coherent and accurate human motion, even in a real-time captured noise scenario.

We also visualize the bird-eye-view of global human translations in Fig. 5. Our method showcases superior performance in predicting both accurate and coherent global human translations, a capability that is particularly evident in handling the complex and challenging motions in FreeMotion-OBJ dataset. In contrast, LiDARCap derives translation estimates from the average locations within point clouds, which can result in inaccuracies and jerkiness due to variations in point distribution, occlusions, and noise interference. Although LiveHPS effectively leverages temporal and spatial data for enhanced accuracy, it falls short of maintaining coherence across its predictions. Our method, on the other hand, achieves both coherent and precise global translations.

5.3 Ablation Study

To prove the superiority of each network module in our LiveHPS++, we conduct ablation study for the network architecture on FreeMotion-OBJ to demonstrate the effectiveness of each module and we also evaluate more details for each module as shown in Tab. 2.

Network Architecture. The network without trajectory-guided body tracker(TBT) yields coherent yet inaccurate results compared to the network without noise-insensitive velocity predictor(NVP) and kinematic-aware pose optimizer(KPO). This discrepancy highlights the critical role of the TBT module in enhancing the accuracy of local motion within noisy environments and the effectiveness of KPO module in optimizing the coherence of sequential motions and translations. Together, the results underscore the significant improvements in both accuracy and coherence brought about by the integration of TBT and KPO in our LiveHPS++'s network architecture.

Table 2: Ablation studies for our network modules on FreeMotion-OBJ. We also evaluate the internal details of each module.

		J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	Jitter↓
Network Module	w/o TBT	71.68/90.08	153.65/164.79	17.52	7.29	33.89
	w/o NVP&KPO	68.37/84.92	134.11/145.01	17.23	7.79	71.82
Trajectory-guided Body Tracker	frame-wise	68.57/86.15	151.73/162.15	16.69	7.54	43.53
	sequence-wise	85.42/106.91	165.71/179.41	19.40	7.10	31.32
Kinematic-aware Pose Optimizer	short-term optimizer	64.04/79.27	132.38/141.57	16.15	7.63	55.50
	long-term optimizer	68.06/83.62	150.90/159.71	16.71	7.21	42.51
Translation Estimation	Average	-	177.13/183.48	-	8.18	94.73
	LIP	-	132.82/140.94	-	8.85	93.56
	LiveHPS	-	130.56/138.71	-	8.12	78.09
Ours		**58.11/72.55**	**128.60/136.94**	**15.85**	**7.01**	**30.96**

Trajectory-guided Body Tracker. We conduct the detailed ablation study on Trajectory-guided Body Tracker(TBT) module by exploring the performance of frame-wise and sequence-wise data normalization. Frame-wise normalization, which does normalization by subtracting point clouds with the average location of each frame, is often utilized in previous LiDAR-based human motion caption methods. It is beneficial for mitigating sensitivity to scale and position variances and network convergence acceleration, achieving accurate prediction but but falls short by omitting essential physical movement information. Sequence-wise normalization, which is sequential normalization by subtracting point clouds with the average location of the first frame, can retain the real-world physical trajectory, achieving smooth but inaccurate results. Our trajectory-guided body tracker facilitates both reduced sensitivity to scale and positional differences, while retaining real-world physical motion information, thus allowing accurate and global human motion results to be obtained.

Kinematic-aware Pose Optimizer. The Kinematic-aware Pose Optimizer (KPO) module refines human joint and translation predictions using velocities predicted by a noise-insensitive velocity predictor. It integrates both short-term and long-term kinematic information for joint-wise optimization. We contrast our method with two others: a short-term approach, which optimizes each frame with the result and the velocity of the previous frame, and a long-term strategy, which optimizes the entire sequence using the results of the first frame and velocity. The short-term optimizer enhances adjacent frame coherence but introduces jerkiness in long sequences and overlooks long-range coherence. Conversely, the long-term optimizer maintains overall coherence but leads to accumulated errors and dependency on the initial frame's accuracy. Our KPO can achieve accurate and coherent results by considering both short-term and long-term kinematic optimization.

Translation Estimation. We further refine human translation estimation, achieving superior accuracy and coherence by minimizing noise impact through trajectory embedding in the TBT module and enhancing translation prediction in the KPO module by leveraging temporal dynamics. This approach significantly surpasses translation estimation methods proposed in LIP and LiveHPS in both acceleration error and jitter metrics, showcasing our method's advanced capability in capturing precise and fluid human motion translations.

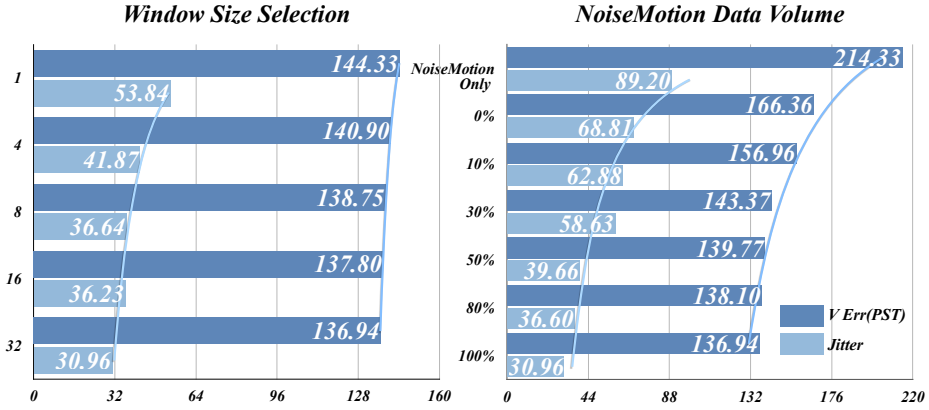

Fig. 6: Ablation study for the temporal window size selection and evaluate the impact of NoiseMotion volume leveraged for training on model performance.

Window Size Selection of KPO. The selection of an optimal temporal window size within the Kinematic-aware Pose Optimizer (KPO) module is crucial for enhancing motion and translation coherence. We experiment with various window sizes 1, 4, 8, 16, and 32. Observing that a window size of 32 yields the best results, as depicted in Figure 6. This configuration led to a slight but consistent decrease in Vertex Error (V Err) and a more pronounced reduction in Jitter, indicating more accurate and coherent sequences. This highlights the significance of a suitable receptive field for optimal sequence coherence and accuracy and confirms our algorithm's adaptability to various temporal lengths.

NoiseMotion Data Volume. We use NoiseMotion to enhance the generalization ability of the network, especially when dealing with noisy data. We discuss the impact of the volume of the synthetic data on network performance in Fig. 6. When we only use NoiseMotion for training, there still exists a domain gap between the real data and synthetic data. We gradually add the NoiseMotion data volume(0%, 10%, 30%, 50%, 80%, and 100%) to real data for training, and the performance gradually improves, which demonstrates the synthetic data is significant for the task.

6 Conclusion

In this paper, we introduce a novel and effective single-LiDAR-based approach, distinguished by its ability to precisely capture accurate and coherent 3D human motions across various unconstrained environments. By fully harnessing the dynamic and kinematic attributes derived from global human movements, we effectively mitigate the adverse effects of significant noise. Additionally, we present a new synthesized motion dataset aimed at augmenting the network's adaptability in noisy conditions. Comprehensive experiments demonstrate the obvious superiority of our method, particularly in terms of local pose accuracy and global pose coherence, rendering our technique highly suitable for practical applications.

References

1. Amin, S., Andriluka, M., Rohrbach, M., Schiele, B.: Multi-view pictorial structures for 3D human pose estimation. In: BMVC (2009). https://doi.org/10.5244/C.27.45
2. Baak, A., Müller, M., Bharaj, G., Seidel, H.P., Theobalt, C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In: ICCV (2011). https://doi.org/10.1109/ICCV.2011.6126356
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: 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. pp. 561–578. Springer (2016)
4. Bregler, C., Malik, J.: Tracking people with twists and exponential maps. In: CVPR (1998). https://doi.org/10.1109/CVPR.1998.698581
5. Burenius, M., Sullivan, J., Carlsson, S.: 3D pictorial structures for multiple view articulated pose estimation. In: CVPR (2013). https://doi.org/10.1109/CVPR.2013.464
6. Cai, Z., Pan, L., Wei, C., Yin, W., Hong, F., Zhang, M., Loy, C.C., Yang, L., Liu, Z.: Pointhps: Cascaded 3d human pose and shape estimation from point clouds. arXiv preprint arXiv:2308.14492 (2023)
7. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
8. Cong, P., Zhu, X., Qiao, F., Ren, Y., Peng, X., Hou, Y., Xu, L., Yang, R., Manocha, D., Ma, Y.: Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. arXiv preprint arXiv:2204.01026 (2022)
9. Cong, P., Zhu, X., Qiao, F., Ren, Y., Peng, X., Hou, Y., Xu, L., Yang, R., Manocha, D., Ma, Y.: Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In: CVPR. pp. 19608–19617 (June 2022)
10. Dai, Y., Lin, Y., Lin, X., Wen, C., Xu, L., Yi, H., Shen, S., Ma, Y., Wang, C.: Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. arXiv preprint arXiv:2303.09095 (2023)
11. De Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. In: ACM SIGGRAPH 2008 papers, pp. 1–10 (2008)
12. Elhayek, A., de Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C.: Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In: CVPR (2015), http://gvv.mpi-inf.mpg.de/projects/convNet_moCap/
13. Guo, K., Taylor, J., Fanello, S., Tagliasacchi, A., Dou, M., Davidson, P., Kowdle, A., Izadi, S.: Twinfusion: High framerate non-rigid fusion through fast correspondence tracking. In: 3DV. pp. 596–605 (2018)
14. Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Transactions on Graphics (TOG) **38**(2), 14:1–14:17 (2019)
15. Habermann, M., Xu, W., Zollhofer, M., Pons-Moll, G., Theobalt, C.: Deepcap: Monocular human performance capture using weak supervision. In: CVPR (June 2020)
16. He, Y., Pang, A., Chen, X., Liang, H., Wu, M., Ma, Y., Xu, L.: Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In: CVPR. pp. 11400–11411 (2021)
17. Holte, M.B., Tran, C., Trivedi, M.M., Moeslund, T.B.: Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. JSTSP **6**(5), 538–552 (2012). https://doi.org/10.1109/JSTSP.2012.2196975

18. Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J.: Towards accurate marker-less human shape and pose estimation over time. In: 3DV. pp. 421–430 (2017). https://doi.org/10.1109/3DV.2017.00055
19. Huang, Y., Kaufmann, M., Aksan, E., Black, M.J., Hilliges, O., Pons-Moll, G.: 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. Jang, D.K., Yang, D., Jang, D.Y., Choi, B., Jin, T., Lee, S.H.: Movin: Real-time motion capture using a single lidar. arXiv preprint arXiv:2309.09314 (2023)
21. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV (2015). https://doi.org/10.1109/ICCV.2015.381
22. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
23. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (June 2019)
24. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: CVPR (June 2020)
25. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019)
26. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: CVPR. pp. 6050–6059 (2017)
27. Li, J., Zhang, J., Wang, Z., Shen, S., Wen, C., Ma, Y., Xu, L., Yu, J., Wang, C.: Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. arXiv preprint arXiv:2203.14698 (2022)
28. Luo, Z., Hachiuma, R., Yuan, Y., Kitani, K.: Dynamics-regulated kinematic policy for egocentric pose estimation. Advances in Neural Information Processing Systems **34** (2021)
29. Noitom Motion Capture Systems. https://www.noitom.com/ (2015)
30. OptiTrack Motion Capture Systems. https://www.optitrack.com/ (2009)
31. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Harvesting multiple views for marker-less 3d human pose annotations. In: CVPR (2017)
32. Peng, X., Zhu, X., Ma, Y.: Cl3d: Unsupervised domain adaptation for cross-lidar 3d detection. AAAI (2023)
33. Ren, Y., Han, X., Zhao, C., Wang, J., Xu, L., Yu, J., Ma, Y.: Livehps: Lidar-based scene-level human pose and shape estimation in free environment. arXiv preprint arXiv:2402.17171 (2024)
34. Ren, Y., Zhao, C., He, Y., Cong, P., Liang, H., Yu, J., Xu, L., Ma, Y.: Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. TVCG (2023)
35. Rhodin, H., Robertini, N., Richardt, C., Seidel, H.P., Theobalt, C.: A versatile scene model with differentiable visibility applied to generative pose estimation. In: ICCV (2015). https://doi.org/10.1109/ICCV.2015.94
36. Robertini, N., Casas, D., Rhodin, H., Seidel, H.P., Theobalt, C.: Model-based outdoor performance capture. In: 3DV (2016), http://gvv.mpi-inf.mpg.de/projects/OutdoorPerfcap/
37. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR (2011)
38. Sigal, L., Bălan, A.O., Black, M.J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV (2010). https://doi.org/10.1007/s11263-009-0273-6

39. Sigal, L., Isard, M., Haussecker, H., Black, M.J.: Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. IJCV **98**(1), 15–48 (2012). https://doi.org/10.1007/s11263-011-0493-4
40. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017)
41. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of Gaussians body model. In: ICCV (2011)
42. Theobalt, C., de Aguiar, E., Stoll, C., Seidel, H.P., Thrun, S.: Performance capture from multi-view video. In: Image and Geometry Processing for 3-D Cinematography, pp. 127–149. Springer (2010)
43. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 109–117 (2017)
44. Vicon Motion Capture Systems. https://www.vicon.com/ (2010)
45. Vlasic, D., Adelsberger, R., Vannucci, G., Barnwell, J., Gross, M., Matusik, W., Popović, J.: Practical motion capture in everyday surroundings. TOG **26**(3), 35–es (2007)
46. Von Marcard, T., Rosenhahn, B., Black, M.J., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In: Computer Graphics Forum. vol. 36, pp. 349–360. Wiley Online Library (2017)
47. Wei, X., Zhang, P., Chai, J.: Accurate realtime full-body motion capture using a single depth camera. SIGGRAPH Asia **31**(6), 188:1–12 (2012)
48. Xsens Technologies B.V. https://www.xsens.com/ (2011)
49. Xu, L., Su, Z., Han, L., Yu, T., Liu, Y., FANG, L.: Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercialrgbd cameras. TPAMI pp. 1–1 (2019)
50. Xu, L., Liu, Y., Cheng, W., Guo, K., Zhou, G., Dai, Q., Fang, L.: Flycap: Markerless motion capture using multiple autonomous flying cameras. TVCG **24**(8), 2284–2297 (Aug 2018)
51. Xu, L., Xu, W., Golyanik, V., Habermann, M., Fang, L., Theobalt, C.: Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In: CVPR (June 2020)
52. Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H.P., Theobalt, C.: Monoperfcap: Human performance capture from monocular video. ACM Transactions on Graphics (TOG) **37**(2), 27:1–27:15 (2018)
53. Xu, Y., Cong, P., Yao, Y., Chen, R., Hou, Y., Zhu, X., He, X., Yu, J., Ma, Y.: Human-centric scene understanding for 3d large-scale scenarios. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20349–20359 (2023)
54. Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., Xu, F.: Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In: CVPR (June 2022)
55. Yi, X., Zhou, Y., Xu, F.: Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. ACM Transactions on Graphics (TOG) **40**(4), 1–13 (2021)
56. Yin, T., Zhou, X., Krähenbühl, P.: Center-based 3d object detection and tracking. CVPR (2021)
57. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. TPAMI (2019)
58. Yuan, Y., Kitani, K.: Residual force control for agile human behavior imitation and extended motion synthesis. Advances in Neural Information Processing Systems **33**, 21763–21774 (2020)
59. Zanfir, A., Bazavan, E.G., Zanfir, M., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Neural descent for visual 3d human pose and shape. arXiv preprint arXiv:2008.06910 (2020)
60. Zhu, X., Ma, Y., Wang, T., Xu, Y., Shi, J., Lin, D.: Ssn: Shape signature networks for multi-class object detection from point clouds. In: ECCV. pp. 581–597. Springer (2020)

61. Zhu, X., Zhou, H., Wang, T., Hong, F., Li, W., Ma, Y., Li, H., Yang, R., Lin, D.: Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. TPAMI (2021)