LiDARCap: Long-range Marker-less 3D Human Motion Capture with LiDAR Point Clouds

Jialian Li[1,*] Jingyi Zhang[1,*] Zhiyong Wang[1] Siqi Shen[1] Chenglu Wen[1] Yuexin Ma[2]
Lan Xu[2] Jingyi Yu[2] Cheng Wang[1,†]
[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University
[2]Shanghai Engineering Research Center of Intelligent Vision and Imaging, ShanghaiTech University
{szljl36,zhangjingyi1,wangzy}@stu.xmu.edu.cn, {siqishen,clwen,cwang}@xmu.edu.cn,
{mayuexin,xulan1,yujingyi}@shangtaitech.edu.cn

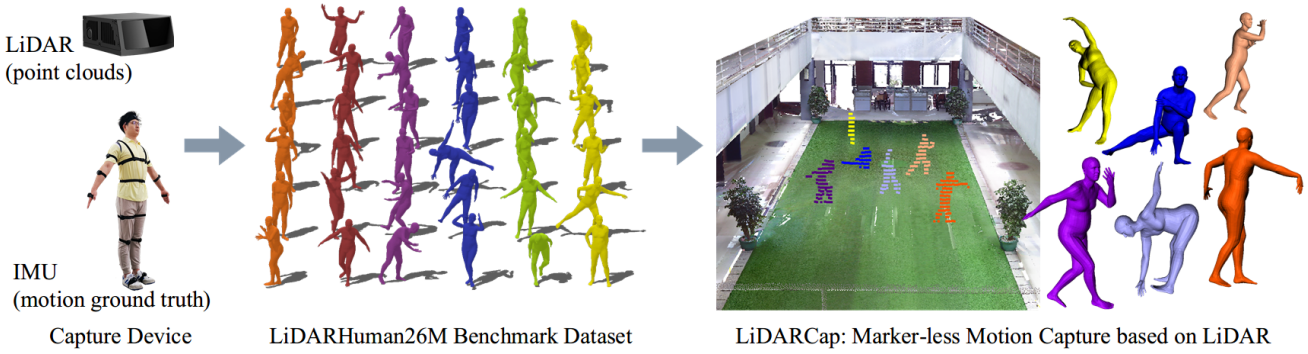


Figure 1. Overview: The proposed LiDARHuman26M benchmark dataset consists of synchronous LiDAR point clouds, RGB images, and ground-truth 3D human motions obtained from professional IMU devices, covering diverse motions and a large capture distance ranging. Based on LiDARHuman26M, we propose LiDARCap, a strong baseline motion capture approach on LiDAR point clouds, which achieves promising results as shown on the right end.

Abstract

Existing motion capture datasets are largely short-range and cannot yet fit the need of long-range applications. We propose LiDARHuman26M, a new human motion capture dataset captured by LiDAR at a much longer range to overcome this limitation. Our dataset also includes the ground truth human motions acquired by the IMU system and the synchronous RGB images. We further present a strong baseline method, LiDARCap, for LiDAR point cloud human motion capture. Specifically, we first utilize PointNet++ to encode features of points and then employ the inverse kinematics solver and SMPL optimizer to regress the pose through aggregating the temporally encoded features hierarchically. Quantitative and qualitative experiments show that our method outperforms the techniques based only on RGB images. Ablation experiments demonstrate that our dataset is challenging and worthy of further research. Finally, the experiments on the KITTI Dataset and the Waymo Open Dataset show that our method can be generalized to different LiDAR sensor settings.

1. Introduction

The past ten years have witnessed a rapid development of marker-less human motion capture [9, 18, 48, 60], with various applications like VR/AR and interactive entertainment. However, conveniently capturing long-range 3D human motions in a large space remains challenging, which is critical for sports and human behavior analysis.

So far, vision-based mocap solutions take the majority in this topic. The high-end solutions require dense optical markers [55, 72] or dense camera rigs [8, 25, 26, 48] for faithfully motion capture, which are infeasible for consumer-level usage. In contrast, monocular capture methods are more practical and attractive. The recent learning-based techniques have enable robust human motion capture from a single RGB stream, using pre-scanned human templates [16, 17, 19, 62, 64] or parametric human models [5, 27, 31, 32, 36, 37, 39]. However, in the long-range capture scenarios where the performers are far away from the cameras, the captured images suffer from degraded and blurred artifacts, leading to fragile motion capture. Vari-

*Equal contribution.
†Corresponding author.

ous methods [65, 66] explore to capture 3D human motions under such degraded and low-resolution images. But such approaches are still fragile to capture the global positions under the long-range setting, especially when handling the textureless clothes or environment lighting changes. In contrast, motion capture using body-worn sensor like Inertial Measurement Units (IMUs) [22, 43, 69] is widely adopted due to its environment-independent property. However, the requirement of body-worn sensors makes them unsuitable to capture motions of people wearing everyday apparel. Moreover, the IMU-based methods will suffer from an accumulated global drifting artifact, especially for the long-range setting. Those motion capture methods [11, 15, 49, 61] using consumer-level RGBD sensors are also infeasible for the long-range capture in a large scene, due to the relatively short effective range (less than 5 m) of RGBD cameras.

In this paper, we propose a rescue to the above problems by using a consumer-level LiDAR. A LiDAR sensor provides accurate depth information of a large-scale scene with a large effective range (up to 30 m). These properties potentially allow capturing human motions under the long-range setting in general lighting conditions, without suffering from the degraded artifacts of visual sensors. Nevertheless, capturing long-range 3D human motions using a single LiDAR is challenging. First, under the long-range setting, the valid observed point clouds corresponding to the target performer is sparse and noisy, making it difficult for robust motion capture. Second, despite the popularity of LiDAR for 3D modeling, most existing work [20, 33, 40, 46, 52, 74] focus on scene understanding and 3D perception. The lack of a large-scale LiDAR-based dataset with accurate 3D human motion annotations leads to the feasibility of a data-driven motion capture pipeline using LiDAR.

To tackle these challenges, we propose *LiDARCap* – the first marker-less, long-range and data-driven motion capture method using a single LiDAR sensor as illustrated in Fig. 1. More specifically, we first introduce a large benchmark dataset *LiDARHuman26M* for LiDAR-based human motion capture. Our dataset consists of various modalities, including synchronous LiDAR point clouds, RGB images and ground-truth 3D human motions obtained from professional IMU-based mocap devices [41]. It covers 20 kinds of daily motions and 13 performers with 184.0k capture frames, resulting in roughly 26 million valid 3D points of the observed performers with a large capture distance ranging from 12 m to 28 m. Note that our LiDARHuman26M dataset is the first of its kind to open up the research direction for data-driven LiDAR-based human motion capture in the long-range setting. The multi-modality of our dataset also brings huge potential for future direction like multi-modal human behavior analysis. Secondly, based on our novel LiDARHuman26M dataset, we provide LiDARCap, a strong baseline motion capture approach on LiDAR point clouds. Finally, we provide a thorough evaluation of various stages in our LiDARCap as well as state-of-the-art image-based methods baselines using our dataset. These evaluations highlight the benefit of the LiDAR-based method against the image-based method under the long-range setting. We also provide preliminary results to indicate that LiDAR-based long-range motion capture remains to be a challenging problem for future investigations of this new research direction. To summarize, our main contributions include:

- We propose the first monocular LiDAR-based approach for marker-less, long-range 3D human motion capture in a data-driven manner.
- We propose a three-stage pipeline consisting of a temporal encoder, an inverse kinematics solver, and an SMPL optimizer to improve pose estimation performance.
- We provide the first large-scale benchmark dataset for LiDAR-based motion capture, with rich modalities and ground-truth annotations. The dataset will be made publicly available.

2. Related Work

Existing Pose Estimation Datasets. In recent years, deep networks have achieved impressive results in inferring the 3D human pose from images or video, and the research focus is tightly intertwined with dataset design. PennAction [71] and PoseTrack [1] are the only ground-truth 2D video datasets, while InstaVariety [28] and Kinetics-400 [6] are pseudo ground truth datasets annotated using a 2D keypoint detector. SURREAL [54] is a large-scale dataset with synthetically-generated but realistic images of people rendered from 3D sequences of human motion capture data. Those datasets have no 3D pose ground truth.

The Human3.6M [23] dataset is a popular benchmark for pose estimation and captured in a controlled indoor environment. It has 3.6 million 3D human poses of 15 activities, and the 3D ground truth is collected using marker-based motion capture systems. Its goal is to predict the 3D locations of 32 joints in the human body defined by SMPL [2]. HumanEva [47] is also restricted to indoor scenarios with static background, providing synchronized video with MoCap. MPI-INF-3DHP [38] is a multi-view dataset captured using a markerless motion capture system in a green screen studio, which records 8 actors performing 8 activities from 14 camera views. Meanwhile, it adopts foreground and background augmentation for addressing the scarcity and limited appearance variability. Another indoor dataset featuring synchronized video, marker-based ground-truth poses, and IMUs called TotalCapture [53] labels for 1.9M

frames. However, those datasets are all collected in indoor areas and have limited variability.

3DPW [57] is an in-the-wild 3D dataset that captures SMPL body poses using IMU sensors and hand-held cameras. It contains 60 video sequences of several outdoor and indoor activities with 7 actors in 18 clothing styles, but 3DPW only provides images without depth information. The PedX [29] collects multi-modal pedestrians data at the large-scale outdoor scenario. Nevertheless, it only provides 3D pseudo label computed using the 2D annotations from a pair of stereo images and LiDAR point cloud. Based on the discussion and practical application requirements above, it is urgent to launch a dataset covering depth information and accurate 3D pose ground truth.

Point Cloud Sequences Processing Methods. Learning-based methods usually process point clouds by considering the Spatio-temporal relationship in point clouds along with time sequences. Choy et al. [7] proposed 4D convolutional neural networks for the spatio-temporal perception that can directly process 3D-videos using high-dimensional convolutions. Huang et al. [21] introduced a spatio-temporal representation learning framework, capable of learning from unlabeled 3D point clouds in a self-supervised fashion. LatticeNet [45] embeds raw point clouds into a sparse permutohedral lattice. Wang et al. [59] proposed a self-supervised schema to learn 4D spatio-temporal features from dynamic point cloud by predicting the temporal order of sampled and shuffled point cloud clips. P4Transformer [12] proposes a Point 4D Transformer to model raw point cloud videos, consisting of a point 4D convolution and a transformer. PSTNet [13] proposes a point spatio-temporal convolution to achieve informative representations of point cloud sequences. Wang et al. [58] proposed anchor-based spatio-temporal attention 3D convolution operations to process dynamic 3D point cloud sequences.

Pose Estimation Methods. As an alternative to the widely used marker-based solutions [4, 51, 56], markerless motion capture [3, 10] technologies alleviate the requirement of body-worn markers and have been widely investigated. EventCap [63] combines model-based optimization and CNN-based human pose detection to capture high-frequency motion details and reduce the drifting in the tracking. Li et al. [34] proposed an approach to volumetric performance capture and novel-view rendering at real-time speed from monocular videos, eliminating the need for expensive multi-view systems or cumbersome pre-acquisition of a personalized template model. RobustFusion [49] proposes a human performance capture system combined with various data-driven visual cues using a single RGBD camera. TailorNet [42] proposes a neural model which predicts clothing deformation in 3D as a function of three factors: pose, shape, and style (garment geometry) while retaining wrinkle detail. Zanfir et al. [70] presented a deep neural network to reconstruct people's 3D pose and shape from an RGB image, including hand gestures and facial expressions. Given a single image and/or a single LiDAR sweep as input, S3 [68] infers shape, skeleton and skinning jointly. However, they focus on the fusion of image and point cloud for human modeling and mainly rely on the synthetic dataset which is composed of the pedestrian behavior, such as walking and running in a short distance.

3. Approach

3.1. Preliminaries

Marker-less 3D motion capture in long-range scenarios is still challenging to the existing methods. As 2D cameras have no depth information, the inherent ambiguity of human joint locations exists in image-based methods, while depth cameras only work in near range. LiDAR sensors have the advantages of both long working range and good distinguish-ability in the depth dimension. In this work, we first develop a human motion dataset containing the LiDAR point clouds on long-range human motion scenarios, together with the synchronized IMU-captured motion ground truth. Our second goal is to establish an end-to-end model that can infer an optimal parametric human model from LiDAR point clouds. We use Skinned Multi-Person Linear Model(SMPL) [2] to represent the pose and shape of a human body compactly. SMPL model contains pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$ associated with human motion, formulated as the relative rotations for 23 joints, to their parent joints and the global body rotation for the root joint, and the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, which control height, weight, limb proportions. The translation parameters $\mathbf{t} \in \mathbb{R}^3$ will be used when the human position is needed. The SMPL model deforms a template triangulated mesh with 6890 vertices based on pose and shape parameters, which is formulated as $\mathbf{V} = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$.

3.2. Dataset: LiDARHuman26M

Long-range motion capture has great potentials in various applications, such as immersive VR/AR experience and action quality assessment. In this paper, we propose the first long-range LiDAR-based motion capture dataset, LiDARHuman26M.

Data Acquisition. We collect data respectively in two scenarios as shown in Fig. 2. The first scene is a patio, which supports far distance human capture. The second scene is an open space between two buildings, supporting a large capturing pitch angle to avoid self-occlusion. The setup details of collection equipment are shown in Tab. 1.

We recruit 13 volunteers (including 11 males and 2 females) to participate in data collection, and they all have signed the consent. The duration for each one varies from 15 to 30 minutes. The distance distribution is shown in

Scene	Range	Height
The scene 1	12-28m	5m
The scene 2	14-24m	7m

Table 1. The setup details of equipment used in two scenes.

Dist(m)	11-13	14-16	17-19	20-22	23-25	26-28
Ratio(%)	0.7	31.4	47.2	17.4	2.4	0.9

Table 2. Distance distribution in the dataset.

Dataset	Frames	Data Source	Long-range?	IMU?	Video?	Real?	Scene
Human3.6M [23]	3.6M	Image	N	Y	Y	Y	Indoor
HumanEva [47]	80.0K	Image	N	Y	N	Y	Indoor
3DPW [57]	51.0K	Image	N	Y	Y	Y	Outdoor
SURREAL [54]	6.5M	Image	N	N	Y	N	Indoor
PedX [29]	10.1K	Point Cloud	Y	N	Y	Y	Outdoor
LiDARHuman26M	184.0K	Point Cloud	Y	Y	Y	Y	Outdoor

Table 3. Statistics and characteristics of related datasets.

Figure 2. Two scenes for data acquisition.

Tab. 2 In summary, LiDARHuman26M provides 184,048 frames, 26,414,383 points, and 20 kinds of daily motions (including walking, swimming, running, phoning, bowing, etc). It consists of three modalities: synchronous LiDAR point clouds, RGB images, and ground-truth 3D human motions from professional IMU-based mocap devices. We preprocessed the data by erasing the background and eliminating the localization error of the IMUs. Details are given in the supplementary materials.

Data Characteristic. Tab. 3 presents statistics of our dataset in comparison to other publicly available 3D human pose datasets. Our LiDARHuman26M dataset has the following features: First, our dataset contains many long-range (up to 28 meters away) human motions, while the image datasets usually have limited capturing distance. Although 3DPW has a certain improvement in this aspect, most of the annotated data still focuses on people nearby. Second, our dataset covers up to 20 daily motions, while HumanEva has only six motions and PedX mainly focuses on walking. Third, our dataset covers three different modalities, including point clouds, RGB videos, and the mocap ground truth provided by IMU. Current image-based datasets do not provide depth information, which is essential for long-range motion capture. SURREAL projects 3D SMPL meshes on the images, and the rendered images are unreal. PedX provides pseudo labels for 3D motions through optimization of LiDAR points along with 2D labels.

Challenge. The long-range characteristic of LiDARHuman26M causes sparsity. As shown in the Fig. 3, the number of points on one person varies greatly, ranging from 30 points to 450 points. Furthermore, it can be manifested in whole body sparsity and partial missing. When the human body moves further from the LiDAR, the points that fall on the body are significantly reduced, resulting in insufficient information to describe the motion. Two different actions may have similar point cloud distributions at low resolution. For example, when capturing at 12m distance, the direction of the human head relative to the body is clear. The data ensures a good alignment between the captured motion and the rough outline. There are only one or two points on the human head at 24m and 27m capturing distance, which is insufficient to confirm the head orientation. Meanwhile, more parts of the body will inevitably miss with the distance increasing. For example, when capturing at 27m distance, the arm is missing, leading to a loss of elbow rotation. The possible reason behind this is body occlusion or too sparse points caused by too far capturing distance.

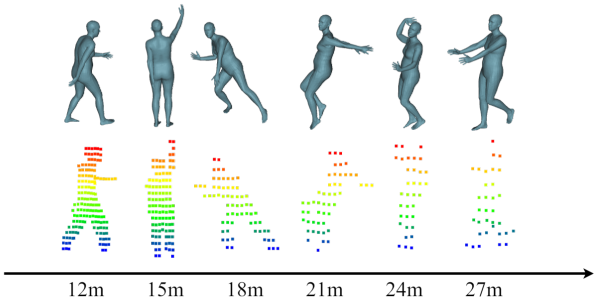


Figure 3. Different human poses with the distance to LiDAR increasing.

3.3. Baseline: LiDARCap

We propose LiDARCap (shown in Fig. 4), a markerless, long-range, and data-driven method for 3D human motion capture using LiDAR point clouds. Trained on LiDARHuman26M, LiDARCap takes point cloud sequences from monocular LiDAR sensor as input and outputs the 3D human motion sequences.

Preprocessing. Given an input LiDAR point cloud sequence $\mathcal{P} = \{\mathbf{P}^{(t)}|t = 1...T\}$ of T frames and each frame contains arbitrary number of points $\mathbf{P}^{(t)} = \{\mathbf{p}_i^{(t)}|i = 1...n_t\}$. We fix the number to 512 by sampling or repeating to perform a unified down-sampling operations.

Temporal Encoder. In this step, we leverage PointNet++ [44] as the backbone to extract a 1024-dim global descriptor $\mathbf{f}^{(t)}$ for each point cloud frame $\mathbf{P}^{(t)}$.

In addition, in order to fuse temporal information, the frame-wise features $\mathbf{f}^{(t)}$ are fed into a two-way GRU (bi-GRU) to generate hidden variables $\mathbf{g}^{(t)}$. At the last of this module, we use $\mathbf{g}^{(t)}$ as input to MLP decoders to predict the corresponding joint locations $\hat{\mathbf{J}}^{(t)} \in \mathbb{R}^{24\times3}$. Here, the loss $\mathcal{L}_{\mathcal{J}}$ of the temporal encoder is formulated as:

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

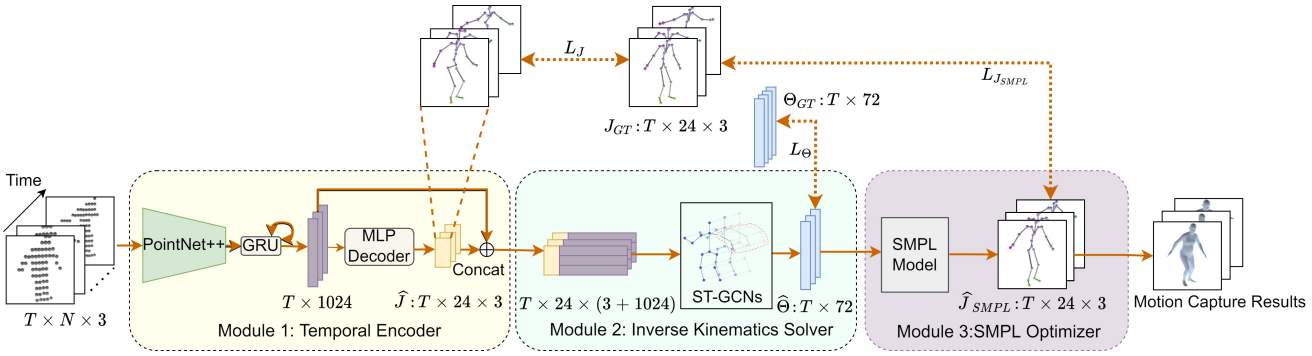


Figure 4. The pipeline of our method with a point cloud sequence as the input consists of a temporal encoder, an inverse kinematic solver, and an SMPL optimizer. T represents the length of the sequence, and N represents the number of points.

where $\mathbf{J}_{GT}^{(t)}$ is the ground truth joint locations of the t-th frame.

Inverse Kinematics Solver. ST-GCN [67] is adopted as the backbone here to extract features of the predicted joints in a graph way. We concatenate the frame-wise global feature with each joint to generate the completed joint features $\mathbf{Q}^{(t)} \in \mathbb{R}^{24\times(3+1024)}$ as the graph node. The output of ST-GCN is subsequently fed into the regressor to compute the joint rotations $\mathbf{R}_{6D}^{(t)} \in \mathbb{R}^{24\times6}$. The 6D rotation is mapped to the final axis-angle format when the loss is computed. We choose the 6D rotation representation as the intermediate results for its better continuity, as demonstrated in [73].

The loss of this module $\mathcal{L}_{\boldsymbol{\Theta}}$ is formulated as:

$$\mathcal{L}_{\boldsymbol{\Theta}} = \sum_t \|\boldsymbol{\theta}_{GT}^{(t)} - \hat{\boldsymbol{\theta}}^{(t)}\|_2^2 \tag{2}$$

where $\boldsymbol{\theta}_{GT}^{(t)}$ is the ground truth pose parameters of the t-th frame .

SMPL Optimizer. We put an SMPL Optimizer module at the last stage to further improve the regression on $\boldsymbol{\theta}$. The joint rotations are fed into an off-the-shelf SMPL model to obtain the 24 joints on the SMPL mesh. $\mathcal{L}_2$ loss between the predicted joints and the ground truth ones is used again in this module to increase the accuracy of the regressed $\boldsymbol{\theta}$ in the last stage. The only difference is that the joints in the first stage are regressed directly through an MLP-based decoder, and here the joints are sampled on the parametric mesh vertices determined by $\boldsymbol{\theta}$.

The loss of this module $\mathcal{L}_{\mathcal{J}_{SMPL}}$ is formulated as:

$$\mathcal{L}_{\mathcal{J}_{SMPL}} = \sum_t \|\mathbf{J}_{GT}^{(t)} - \hat{\mathbf{J}}_{SMPL}^{(t)}\|_2^2 \tag{3}$$

where $\mathbf{J}_{SMPL}^{(t)}$ is the joint locations sampled from the SMPL mesh parameterized by the pose parameter $\hat{\boldsymbol{\theta}}^{(t)}$.

This step provides stronger constraints on the regression of $\boldsymbol{\theta}$ in a geometrically intuitive way. The ablation experiment is conducted to demonstrate its necessity, and more details can be seen in Sec. 4.2.

To sum up, our pipeline can be trained through optimizing the united loss function $\mathcal{L}$ formulated as below in an end-to-end way:

$$\mathcal{L} = \mathcal{L}_{\mathcal{J}} + \mathcal{L}_{\boldsymbol{\Theta}} + \mathcal{L}_{\mathcal{J}_{SMPL}} \tag{4}$$

Training details. We train our method for 200 epochs with Adam optimizer [30] and set the dropout ratio as 0.5 for the GRU layers and ST-GCN module. We apply batch normalization layer after every convolutional layer except the final output layer before the decoder. During training, one NVIDIA GeForce RTX 3090 Graphics Card is utilized. The batch size is set to be 8, while the learning rate is set to be 1×10^{-4}. The decay rate is 1×10^{-4}. The network architecture involved in the evaluation section is trained using the most suitable learning rate until convergence. We train our method on the proposed LiDARHuman26M dataset, and experiment details are provided in Sec. 4.

4. Experiments

4.1. Comparison

The proposed LiDARCap method performs well in predicting human motions in long-range scenarios, as shown in Fig. 5. For further investigation, our method was compared with the state-of-the-art (SOTA) image-based motion capture methods. Quantitative and qualitative comparisons with HMR [27] and VIBE [31] are conducted where the latter also relies on the temporal encoding.

As shown in the Fig. 6, benefited from the 3D distinguish-ability of the LiDAR point clouds, our method outperforms the image-based methods. The performance of HMR is badly contaminated by the low quality of the distant images, while VIBE can speculate some unclear motions with the help of sequential constrain.

Tab. 4 shows the corresponding quantitative comparisons using different evaluation metrics. We report Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE), Mean Per Joint Position Error (MPJPE), Percentage of Correct Keypoints (PCK), and Per Vertex Error (PVE). The error

Figure 5. 3D capturing results on long-range human motions. For each body motion, the top row shows the reference images, the middle row shows the input LiDAR points, and the bottom row shows the captured motion results on the LiDAR view.

metrics are measured in millimeters. In addition, Acceleration error(m/s^2) is also recorded as an important evaluation indicator for sequence data. Benefiting from the effective use of 3D spatial information, our method significantly outperforms HMR and VIBE.

4.2. Evaluation

To study the effect of different components of our method, we conduct two ablation experiments. The first experiment validates the effectiveness of the combination of PointNet++ and ST-GCN. The second one verifies the effectiveness of the combination of the inverse kinematics solver and SMPL optimizer.

Evaluation on network structure. For simplicity, our method is called P++/ST-GCN. On the one hand, we replace the PointNet++(P++) backbone with other diligently-designed network structures. They are P4Transformer(P4T) [12], attention(ATT) module in [24] and the voting(VOT) module in [35]. In order to fuse the spatial-temporal information, we use the P4Transformer instead of the original one as the backbone of the latter two. On the other hand, we need to evaluate whether it is necessary to leverage ST-GCN to exploit the joint features over the temporal dimension instead of biGRU. Tab. 5 shows the comparison results mentioned above, from which we find that on our dataset, the more complicated operation like attention and voting will result in a decrease in the performance. Global features

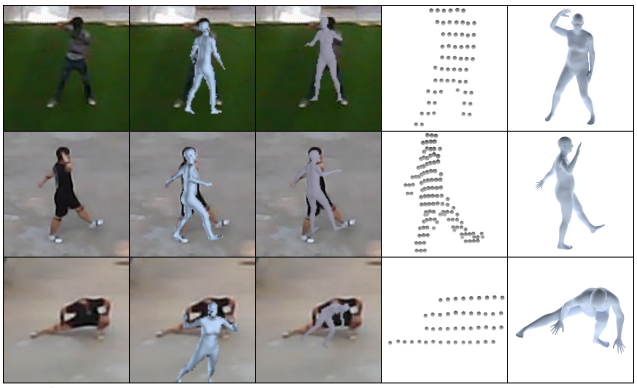


Figure 6. Qualitative comparison with the image-based methods. Our method provides accurate human pose, while the results of SOTA image-based methods contain large errors.

Method	MPJPE↓	PA-MPJPE↓	PCK0.5↑	PCK0.3↑	Accel↓	PVE↓
HMR	224.86	130.71	0.67	0.49	22.07	284.15
VIBE	154.61	108.19	0.82	0.64	12.49	191.55
Ours	**79.31**	**66.72**	**0.95**	**0.86**	**4.52**	**101.64**

Table 4. Quantitative comparison of our method and image-based methods in terms of capturing accuracy.

Method	MPJPE↓	PA-MPJPE↓	PCK0.5↑	PCK0.3↑	Accel↓	PVE↓
P++/GRU	86.43	72.19	0.94	0.83	5.20	109.48
ATT/ST-GCN	96.28	75.21	0.92	0.81	4.66	120.42
P4T/ST-GCN	79.52	**66.25**	**0.95**	**0.86**	4.54	101.77
VOT/ST-GCN	146.20	100.40	0.83	0.67	7.10	185.33
P++/ST-GCN(Ours)	**79.31**	66.72	**0.95**	**0.86**	**4.52**	**101.64**

Table 5. Quantitative evaluation of different encoders and sequential processing methods.

Method	MPJPE↓	PA-MPJPE↓	PCK0.5↑	PCK0.3↑	Accel↓	PVE↓
P++ w/o ST-GCN	85.93	70.61	0.94	0.84	4.91	109.27
P++/ST-GCN w/o $\mathcal{L}_{J_{SMPL}}$	87.13	69.35	0.94	0.83	4.98	110.23
P++/ST-GCN(Ours)	**79.31**	**66.72**	**0.95**	**0.86**	**4.52**	**101.64**

Table 6. Quantitative evaluation of different combinations of stages.

help achieve the best performance, and there is no significant difference between P++ and P4T. Moreover, introducing the kinematic tree can help localize the adjacent joints better than the biGRU, which can only impact the frame-wise global features. The convolution on the same joints over the time step also ensures the continuity and consistency explicitly.

Evaluation on stages. The necessity of all the three modules proposed in Sec. 3.3 is demonstrated in Tab. 6. Among the three modules, the temporal encoder used to extract point features is indispensable. The performance difference is that the joint locations help the network learn the motion features more efficiently both in the first and the third stage. Among them, the previous one is used to constrain the solution domain of the rotations, while the latter one serves an important role as a posteriori validation.

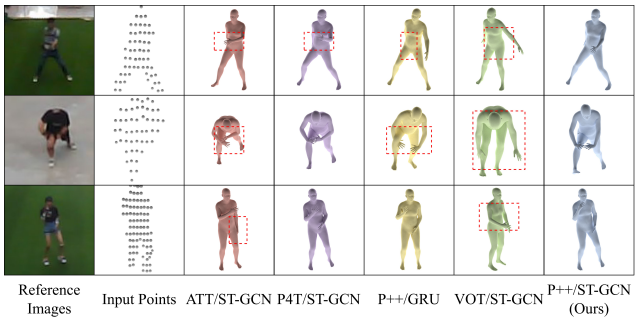


Figure 7. Qualitative results of different network structures. P4T/ST-GCN and our method can capture more consistent motions than other methods.

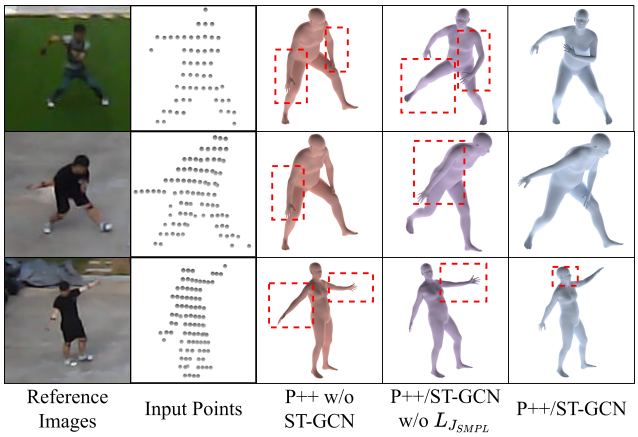
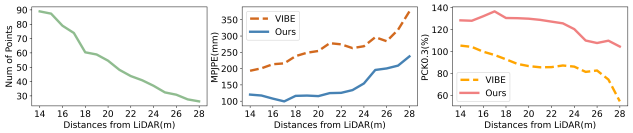


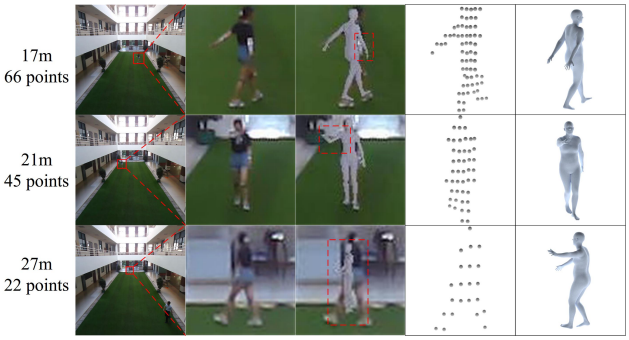
Figure 8. Qualitative results of different combination of stages. Other two methods cannot predict accurate human motions in some cases for lacking enough constraint.

4.3. Distance Analysis

Benefiting from the excellent characteristics of the point cloud, our method has achieved good results in long-range motion capture. However, the latest version of the algorithm still cannot do well at a too far distance. We take the sequence with the most dramatic distance change as the illustration, in which the trajectory of the volunteer ranges from 15 to 28 meters from the LiDAR. As can be seen from Fig. 9a, the number of points projected on a person decreases sharply as the distance increases, and at the farthest distance, the number is less than 30. At this moment, apart from the outline of the human, it is difficult to discriminate the detailed movements. The low resolution of long-range images also poses great challenges to image-based ones. Although the performance of our method and VIBE decreases as the distance increases, we can still maintain better results which is shown in Fig. 9b. This is because, with the distance increases, the human can still be segmented clearly from the background in the point cloud while the image pixels of human is more prone to mix with the background ones, though the resolution of both data sources is decreasing.



(a) The number of points and performance of VIBE and ours over the distance.



(b) The qualitative results over the chosen three distances. The columns from left to right are the original images, the enlarged images, VIBE results, the point clouds, and our results.

Figure 9. Evaluation of VIBE and ours over the different distances. The figures show that performance will decrease as distance increases. However, our algorithm can still achieve more convincing results than VIBE.

4.4. Results on KITTI and Waymo Dataset

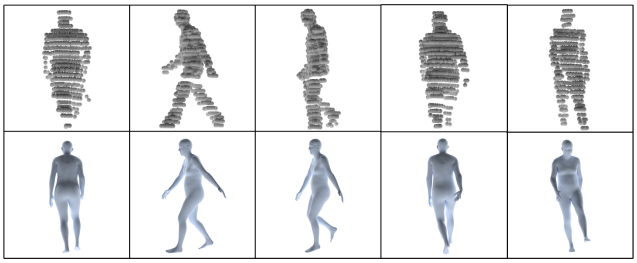
In order to verify the generalization of our pipeline, we test our method on point cloud sequences of pedestrians from the KITTI Detection Dataset [14] and the Waymo Open Dataset [50]. Fig. 10 shows some qualitative results.

It can be seen that our algorithm can learn the correct footsteps and global orientation of pedestrians. For the clear part of the upper limb, our method can make the correct placement, while for the ambiguous one, it will make reasonable guesses through prior information of time series.

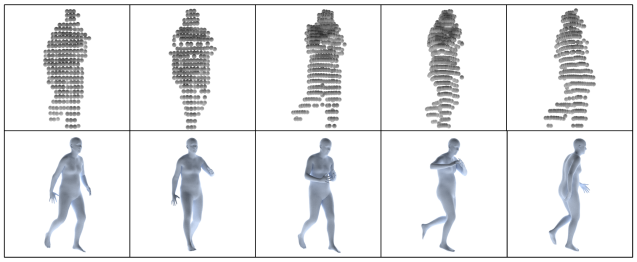
5. Discussion

Limitation. First, the scenario in LiDARHuman26M is flat, open, and unobstructed, which is too idealistic compared to the real applications. Second, the shape parameters β and more complex scenes with occlusions and interactions of multi-person are lacked in the dataset LiDARHuman26M. Third, the proposed baseline LiDARCap method is not robust enough to handle varying density of point clouds from different distances and devices. Accurate human motion capture on sparse LiDAR point clouds is still an open challenge.

Conclusion. We present LiDARHuman26M, the first of its kind dataset to open up the research direction of data-driven LiDAR-based human motion capture in the long-range setting. LiDARHuman26M consists of various modalities, including synchronous LiDAR point clouds, RGB images, and ground-truth 3D human motions obtained from professional IMU devices. It covers 20 kinds of daily motions and 13 performers with 184.0k capture frames, with a large capture distance ranging from 12 m to 28 m. Based on LiDARHuman26M, we propose a strong baseline method, LiDARCap, the first marker-less, long-range, and data-driven human motion capture method for monocular LiDAR sensor. Specifically, the proposed LiDARCap extracts global features of LiDAR point clouds. It then employs the inverse kinematics solver and SMPL optimizer to hierarchically regress the human pose through aggregating the temporally encoded features. Quantitative and qualitative experiments show that our method outperforms the methods based only on RGB images. The experiments on the LiDAR data of the KITTI Dataset and the Waymo Open Dataset show that our method can be generalized to different LiDAR sensor settings.



(a) Qualitative results of the KITTI Dataset.



(b) Qualitative results of the Waymo Open Dataset.

(c) Qualitative results of two sequences from the Waymo Open Dataset.

Figure 10. Qualitative results on autonomous driving datasets. Our method can discriminate the correct motions of the clear parts of the point cloud and give a reasonable guess of the invisible ones.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Anton Milan, Eldar Insafutdinov, Leonid Pishchulin, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018. 2

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 2, 3

[3] Christoph Bregler and Jitendra Malik. Tracking people with twists and exponential maps. *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, pages 8–15, 1998. 3

[4] Xsens Technologies B.V. https://www.xsens.com/, 2019. 3

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[6] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 2

[7] Christopher Bongsoo Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3070–3079, 2019. 3

[8] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):69, 2015. 1

[9] Andrew J. Davison, Jonathan Deutscher, and Ian D. Reid. Markerless motion capture of complex full-body movement for character animation. In *Eurographics Workshop on Computer Animation and Simulation*, 2001. 1

[10] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *ACM SIGGRAPH 2008 papers*, 2008. 3

[11] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. Fusion4D: Real-time Performance Capture of Challenging Scenes. In *ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques*, 2016. 2

[12] Hehe Fan, Yi Yang, and Mohan S. Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14199–14208, 2021. 3, 6

[13] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and M. Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. In *ICLR*, 2021. 3

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 8

[15] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo and motion reconstruction using a single rgbd camera. *ACM Transactions on Graphics (TOG)*, 2017. 2

[16] 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. 1

[17] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[18] Nils Hasler, Bodo Rosenhahn, Thorsten Thormahlen, Michael Wand, Juergen Gall, and Hans-Peter Seidel. Markerless motion capture with unsynchronized moving cameras. In *Computer Vision and Pattern Recognition (CVPR)*, pages 224–231, 2009. 1

[19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11400–11411, 2021. 1

[20] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11108–11117, 2020. 2

[21] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6515–6525, 2021. 3

[22] 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

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:1325–1339, 2014. 2, 4

[24] Haiyong Jiang, Jianfei Cai, and Jianmin Zheng. Skeleton-aware 3d human shape reconstruction from point clouds. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5430–5440, 2019. 6

[25] 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 *Proceedings of the IEEE Inter-*

national Conference on Computer Vision, pages 3334–3342, 2015. 1

[26] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[27] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 1, 5

[28] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5607–5616, 2019. 2

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

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 5

[31] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 5

[32] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 1

[33] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12697–12705, 2019. 2

[34] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *ECCV*, 2020. 3

[35] Shile Li and Dongheui Lee. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11919–11928, 2019. 6

[36] 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, Oct. 2015. 1

[37] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[38] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal V. Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 2

[39] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4), 2017. 1

[40] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet ++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220, 2019. 2

[41] INC. NOITOM INTERNATIONAL. https://noitom.com/, 2021. 2

[42] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7363–7373, 2020. 3

[43] Gerard Pons-Moll, Andreas Baak, Juergen Gall, Laura Leal-Taixe, Meinard Mueller, Hans-Peter Seidel, and Bodo Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *2011 International Conference on Computer Vision*, pages 1243–1250. IEEE, 2011. 2

[44] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space, 2017. 4

[45] Radu Alexandru Rosu, Peer Schutt, Jan Quenzel, and Sven Behnke. Latticenet: Fast spatio-temporal point cloud segmentation using permutohedral lattices. *Autonomous Robots*, pages 1–16, 2021. 3

[46] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–779, 2019. 2

[47] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87:4–27, 2009. 2, 4

[48] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of Gaussians body model. In *International Conference on Computer Vision (ICCV)*, 2011. 1

[49] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *ECCV*, 2020. 2, 3

[50] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott M. Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2451, 2020. 8

[51] Vicon Motion Systems. https://www.vicon.com/, 2019. 3

[52] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6411–6420, 2019. 2

[53] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John P. Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *BMVC*, 2017. 2

[54] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017. 2, 4

[55] Vicon Motion Systems. https://www.vicon.com/, 2019. 1

[56] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John C. Barnwell, Markus H. Gross, Wojciech Matusik, and Jovan Popović. Practical motion capture in everyday surroundings. *ACM SIGGRAPH 2007 papers*, 2007. 3

[57] 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*, 2018. 3, 4

[58] Guangming Wang, Hanwen Liu, Muyao Chen, Yehui Yang, Zhe Liu, and Hesheng Wang. Anchor-based spatio-temporal attention 3-d convolutional networks for dynamic 3-d point cloud sequences. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2021. 3

[59] Haiyan Wang, Liang Yang, Xuejian Rong, Jinglun Feng, and Yingli Tian. Self-supervised 4d spatio-temporal feature learning via order prediction of sequential point cloud clips. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3761–3770, 2021. 3

[60] Yangang Wang, Yebin Liu, Xin Tong, Qionghai Dai, and Ping Tan. Outdoor markerless motion capture with sparse handheld video cameras. *Transactions on Visualization and Computer Graphics (TVCG)*, 2017. 1

[61] 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

[62] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[63] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Ming Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4967–4977, 2020. 3

[64] 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. 1

[65] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A Jeni, and Fernando De la Torre. 3d human shape and pose from a single low-resolution image with self-supervised learning. In *ECCV*, 2020. 2

[66] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A Jeni, and Fernando De la Torre. 3d human pose, shape and texture from low-resolution images and videos. *TPAMI*, 2021. 2

[67] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 5

[68] Ze Yang, Shenlong Wang, Siva Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *CVPR*, 2021. 3

[69] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Trans. Graph.*, 40:86:1–86:13, 2021. 2

[70] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3d human pose and shape. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14479–14488, 2021. 3

[71] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. *2013 IEEE International Conference on Computer Vision*, pages 2248–2255, 2013. 2

[72] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 1

[73] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, 2019. 5

[74] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018. 2