Neighborhood-Enhanced 3D Human Pose Estimation with Monocular LiDAR in Long-Range Outdoor Scenes

Jingyi Zhang[1,2], Qihong Mao[1,2], Guosheng Hu[3], Siqi Shen[1,2], Cheng Wang[1,2*]

[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, China
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University, China
[3] Oosto, Belfast, UK
{zhangjingyi1, maoqihong}@stu.xmu.edu.cn; {siqishen, cwang}@xmu.edu.cn; huguosheng100@gmail.com

Abstract

3D human pose estimation (3HPE) in large-scale outdoor scenes using commercial LiDAR has attracted significant attention due to its potential for real-life applications. However, existing LiDAR-based methods for 3HPE primarily rely on recovering 3D human poses from individual point clouds, and the coherence cues present in the neighborhood are not sufficiently harnessed. In this work, we explore spatial and contexture coherence cues contained in the neighborhood that leads to great performance improvements in 3HPE. Specifically, firstly, we deeply investigate the 3D neighbor in the background (3BN) which serves as a spatial coherence cue for inferring reliable motion since it provides physical laws to limit motion targets. Secondly, we introduce a novel 3D scanning neighbor (3SN) generated during the data collection and 3SN implies structural edge coherence cues. We use 3SN to overcome the degradation of performance and data quality caused by the sparsity-varying properties of LiDAR point clouds. In order to effectively model the complementation between these distinct cues and build consistent temporal relationships across human motions, we propose a new transformer-based module called the CoherenceFuse module. Extensive experiments conducted on publicly available datasets, namely LidarHuman26M, CIMI4D, SLOPER4D and Waymo Open Dataset v2.0, showcase the superiority and effectiveness of our proposed method. In particular, when compared with LidarCap on the LidarHuman26M dataset, our method demonstrates a reduction of 7.08mm in the average MPJPE metric, along with a decrease of 16.55mm in the MPJPE metric for distances exceeding 25 meters. The code and models are available at *https://github.com/jingyi-zhang/Neighborhood-enhanced-LidarCap*.

Introduction

3D human pose estimation (3HPE) in unconstrained environments is a rapidly advancing field with lots of promising research work (Kim et al. 2022; Joo, Neverova, and Vedaldi 2020; Zhang et al. 2021; Jin et al. 2020; Wang et al. 2022; Zhan et al. 2022). However, accurate 3HPE in long-range outdoor scenes, which has diverse and impactful applications in action recognition, sports analysis, AR/VR, autonomous driving, etc, remains challenging.

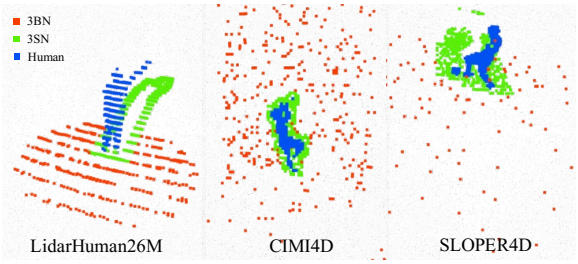

Figure 1: The visualization of 3SN and 3BN. Our method treats 3SN sequence, 3BN sequence, and human sequence as input to predict 3D human motions.

To effectively address long-range 3HPE, various modalities have distinctive challenges that require specific solutions. Some researchers (Xu et al. 2021, 2020b; Wan et al. 2021) focus on recovering 3D human motions from degraded and low-resolution images, due to factors such as poor lighting and long camera-subject distances. Other researchers (Zhang et al. 2022b; Sun et al. 2020; Li et al. 2020; Zhang et al. 2022a) concentrate on alleviating the influence of ill-posed to global location prediction from RGB images since the camera is unable to provide accurate depth information. RGBD-based methods are also unsuitable for long-range 3HPE, due to the limited effective range (less than 5 m). In contrast, body-worn sensors like Inertial Measurement Units (IMUs) have environment-independent properties, IMUs-based methods (Yi et al. 2022; Yi, Zhou, and Xu 2021b; Marcard et al. 2017; Huang et al. 2018b) dedicated to achieving convenient 3HPE by minimizing the number of wearable devices. HSC4D (Dai et al. 2022) employs LiDAR sensors to obtain depth information and correct an accumulated global drifting artifact that occurs in IMUs when working in long-range outdoor scenes.

Despite the success of these sensors, benefitting from the LiDAR's inherent insensitivity to lighting conditions, its capacity to acquire precise depth information, and its extended detection range, the LiDAR-based method is the preferred choice for conducting motion capture in daily scenarios characterized by extensive distances and expansive environments. Recently, researchers started exploring the *LiDAR sensor* in human motion capture. LidarCap(Li et al. 2022a) achieves remarkable results in 3D human motion capture within a range of 30 meters, relying solely on individual human point clouds

*Corresponding author.

obtained from a single monocular LiDAR sensor, demonstrating its outstanding potential. Multimodal fusion methods (Ren et al. 2022; Fürst et al. 2020; Kim et al. 2019) improve algorithm robustness and accuracy by combining LiDAR data with data from other modalities. Nevertheless, previous methods overlook the fact that LiDAR sensors capture data not only from humans but also from the environment. We hypothesize that environmental information, which correlates with human behavior, is a key element for overcoming fragile motion capture caused by sparsity. SLOPER4D (Dai et al. 2023) and CIMI4D (Yan et al. 2023) incorporate environmental information in the post-optimization stage, they utilize environmental information as a physical constraint to enhance the initial human pose collected by a motion capture system. However, the inclusion of additional physical constraint terms leads to a significant increase in computation time. Therefore, these methods are not suitable for real-time applications.

To achieve *efficient* and accurate LiDAR-based 3HPE, in this paper, we utilize spatial and contexture coherence cues about human behavior in the neighborhood to enhance the performance of the LiDAR-based 3HPE method. Specifically, we introduce two types of coherence cues in the neighborhood.

Firstly, the 3D neighbor in the background (3BN) is extracted from surrounding scenes and preserves the accurate distribution of the real world. The spatial coherence cues implied in 3BN are important for 3HPE, which are not easily accessible in images or IMUs data. It provides a physical law and spatial priority to motion targets. Therefore, when human motion is unreliable, the target poses will be extrapolated from spatial coherence cues contained in 3BN.

Secondly, the 3D scanning neighbor (3SN) is defined based on the characteristics of LiDAR point clouds. As the laser beam scatters to a certain size, the portion of the beam that illuminates the edges of the human body continues to travel until it hits more distant objects (Robosense 2023). Hence, the 3SN provides contexture coherence cues related to the transition from the human structure to the background. By utilizing this additional information, our approach can address the issue of fragile motion capture caused by degraded LiDAR point clouds at larger capturing distances.

The introduction of 3SN originally aimed at mitigating point cloud degradation in *distant* regions. We observed that projecting the 3SN onto the plane proved more advantageous when the human point cloud is densely populated in *close* proximity. This observation led us to speculate that 3SN encapsulates not only structural edge coherence cues but also vital human motion cues. In light of this, we devised a self-attention mechanism tailored to 3SN. Furthermore, recognizing the potential synergy among the spatial coherence cues in 3BN, the abundant coherence cues within 3SN, and the motion cues within human points, we introduced a cross-attention structure to harmoniously fuse these diverse cues. The whole transformer-based module is named the "CoherenceFuse" module.

To this end, by combining the 3SN with 3BN, our method offers a more comprehensive and reliable way to understand and predict human behavior in long-range outdoor scenes. Despite the simplicity of our method, it outperforms the SOTA learning-based and optimized-based methods on 4 public datasets without the need for additional modal data or physical constraint terms. Specifically, our method compared with the baseline method LidarCap (Li et al. 2022a) on the LidarHuman26M dataset (Li et al. 2022a) reduces the mean per joint position error (MPJPE) metric by 7.08mm and improves the Percentage of Correct Keypoints (PCK30) metric by 1.94%. In the case of distant targets (>25m), our method achieves even greater improvements, with the MPJPE metric reducing by around 16.55mm and the PCK30 metric improving by 5.64%. These results demonstrate the effectiveness of our approach in utilizing the coherence cues contained in the neighborhood to accurately predict and capture human motion, even in challenging environments. To summarize, this work has the following key contributions:

- We deeply investigate 3D neighbor in the background (3BN), which effectively leverages spatial coherence cues to predict reliable 3D human motion. In addition, we propose a 3D scanning neighbor (3SN), which creatively employs the contexture coherence cues to compensate for the data degradation caused by increasing distance, enabling accurate estimation of human motion even at long ranges.
- We introduce a CoherenceFuse module to build consistent temporal relationships across human motions and efficiently integrate the information encompassed by 3BN, 3SN, and human point clouds. Thus, CoherenceFuse can enable these diverse cues to complement each other.
- Our method by fully utilizing spatial and structure coherence cues contained in the neighborhood significantly outperforms the baseline method in both close and distant ranges. It offers a simple yet effective way to perform 3D LiDAR-based 3HPE.

Related Work

The past ten years have witnessed a rapid development of 3HPE. Inertial methods (Huang et al. 2018a; Yi, Zhou, and Xu 2021b; Yi et al. 2022) use Inertial Measurement Units (IMUs) to recover human motion with environment-independent properties. Image-based methods (Xie, Bhatnagar, and Pons-Moll 2023; Zanfir, Marinoiu, and Sminchisescu 2018; Xu et al. 2018, 2020a; Habermann et al. 2020; Kocabas, Athanasiou, and Black 2020) reconstruct 3D humans from images, those methods are more practical and attractive when applied in sufficient light and moderate distance condition. RGBD-based methods (Su et al. 2020, 2021; Bhatnagar et al. 2022) using RGBD sensors are feasible for short-range human motion capture. Since we target 3HPE in long-range outdoor scenes with a monocular LiDAR, we review previous works that are most related to our method.

LiDAR-Based Methods for 3HPE

The growing interest in the convenient capture of human motions under long-range scenario settings has led to the growing popularity of LiDAR-based motion capture methods. LidarCap (Li et al. 2022a) leverages point clouds of the human body collected by a single static LiDAR sensor within a range of 30 meters to predict corresponding human

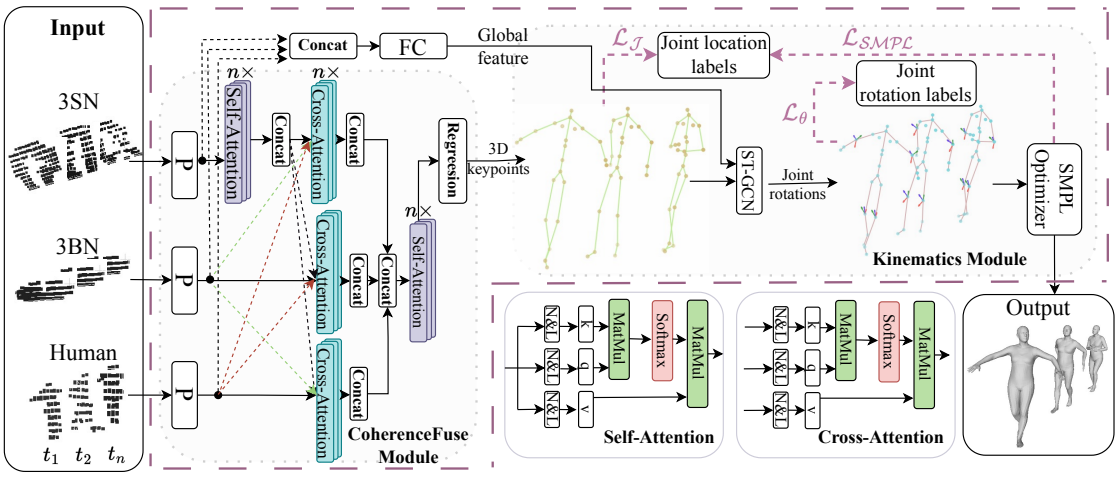

Figure 2: **Overall pipeline.** Given input point clouds of the 3BN, the 3SN and an individual human with XYZ coordinates, PointNet++ (Qi et al. 2017) extracts features separately, and then those features are aggregated by a fully connected layer. Multi-head self-attention layers and multi-head cross-attention layers are utilized to predict joints' location, where n=8. ST-GCN (Yan, Xiong, and Lin 2018) is utilized to predict joints' rotation. At last, SMPL (Bogo et al. 2016) template corrects the final joints' location.

motions. LIP (Ren et al. 2022) extracts rough global human pose from point clouds, then utilizes IMUs to refine local dynamic motions. In addition, there are many LiDAR-camera-based sensor fusion methods for 3D HPE. FusionPose (Cong et al. 2022) exploits the inherent geometry constraints of point clouds and 2D keypoints on images for self-supervision. HPERL (Fürst et al. 2020) integrates features of images and point clouds for superior precision. Zheng et al. (Zheng et al. 2021) utilize a 2D keypoints heatmap predicted from 2D images to augment the point clouds. Although those methods have shown excellent performance in long-range 3D human motion capture, they only utilize the collected data of the human body and ignore coherence cues contained in the environment, which we assume is critical information to estimate human motion when human body data is degraded severely.

Scene-Aware 3D Human Motion Capture

Many existing algorithms resort to multi-stage optimization to estimate global human pose and human-scene interaction. SLOPER4D (Dai et al. 2023) provides reconstructed scene point clouds, and they use scene geometry with several physic-based terms to perform joint optimization. CIMI4D (Yan et al. 2023) concentrates on off-grounding action and facilitates a detailed exploration of human-scene interaction by using a blending optimization process. However, those multi-stage optimization methods are inappropriate for time-critical applications. Luo et al. (Luo et al. 2022) propose a one-stage embodied scene-aware 3HPE method based on a simulated agent's proprioception and scene awareness, along with external third-person observations. PORX (Hassan et al. 2019) formulates the inter-penetration constraint and concat constraint to make use of the 3D scene information. Nonetheless, those methods need to obtain a prescanned environment, which is unsuitable for the large-scale scenario.

Transformer-Based 3D Human Motion Capture

There is already much research about transformer-based methods for estimating 3D human motion (Xu et al. 2022; Yi, Zhou, and Xu 2021a). LPFormer (Ye et al. 2023) employ many blocks of a multi-head self-attention to regress 3D human joints location from LiDAR point clouds by fusing points voxel features, points features, and box features. MHFormer (Li et al. 2021) presented to relieve an inverse problem where multiple feasible solutions exist. It utilizes self-attention to capture relationships across solutions features, and then cross-attention is applied to aggregate the multi-hypothesis features and predict the final 3D pose.

Methodology

Our task is to estimate human motion using monocular LiDAR. The input of our method is sequential point clouds and the output is the 3D human motion sequences in terms of joint angles, global joint locations, and global rotation. The overall pipeline is shown in Fig.2, which incorporates two modules. (1) the CoherenceFuse module: a transformer-based feature fusion module, which incorporates the 3D scanning neighbor (3SN) and the 3D neighbor in the background (3BN) into the 3HPE network. The 3SN and 3BN shown in Fig. 1. (2) the kinematics module: a neural motion estimator. With the purpose of providing more convincing evidence to prove the benefits brought by incorporating 3SN and 3BN, the structure of the kinematics module follows LidarCap (Li et al. 2022a), which includes ST-GCN (Yan, Xiong, and Lin 2018) and SMPL optimizer.

3D Neighbor in the Background (3BN)

Given individual human LiDAR point clouds at t frame $p_{ht} = \{p_{1t}, p_{2t}, ..., p_{nt}\}$, we calculate the global position of the human p_{ct} by $p_{ct} = \frac{1}{n} * \sum p_{it}$, where i=1 to n. Any point in the environment within a two-meter radius of p_{ct} is

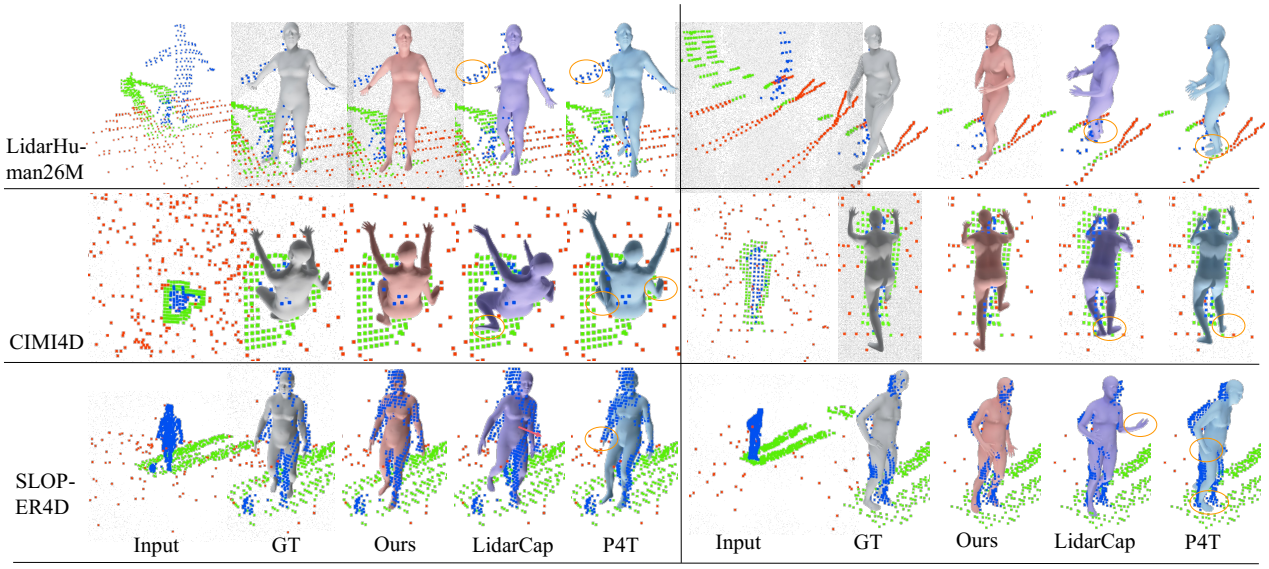

Figure 3: Quanlitive results on 3 public datasets. The blue points mean human points. The red points mean 3BN. The green points mean 3SN.

Method \ Dataset	LidarHuman26M					CIMI4D					SLOPER4D				
Metric	MPJPE	PA-	PCK3	PCK5	AE	MPJPE	PA-	PCK3	PCK5	AE	MPJPE	PA-	PCK3	PCK5	AE
LidarCap	79.31	66.72	86.0	95.0	45.2	130.43	97.95	70.22	87.44	31.11	101.89	78.93	78.15	89.77	40.09
HSC2	79.48	68.9	85.33	94.35	**29.49**	-	-	-	-	-	-	-	-	-	-
P4T	127.77	98.44	72.58	85.64	151.67	146.93	108.94	65.26	83.96	97.13	103.56	79.34	77.42	90.04	66.62
CLIFF	80.55	**60.03**	86.03	94.64	92.48	-	-	-	-	-	-	-	-	-	-
Ours	**72.23**	61.67	**87.94**	**95.79**	38.28	**121.77**	**93.15**	**72.81**	**89.03**	**25.81**	**96.80**	**76.70**	**79.22**	**90.51**	**38.55**

Table 1: Comparison results on LidarHuman26M, CIMI4D and SLOPER4D with learning-based method LidarCap, P4T and optimized-based method HSC2.

defined as the 3BN (p_{bt}). The two-meter radius parameter is defined based on the human arm span being approximately 1 meter. Therefore, the area within a 1-meter radius of p_{ct} is the interaction region. The spatial coherence cues contained in this region can provide physical references for 3HPE, such as the positions of foot and palm joints. The area, which is outside the 1-meter radius but within the 2-meter radius, is where interaction with the human is about to occur. The spatial coherence cues in this region provide interaction priors for 3HPE and constrain motion targets.

3D Scanning Neighbor (3SN)

Given a set of 3D points p_{ht} representing a human body in Cartesian coordinates at t frame, we can transform each point $p_{it} = (x_{it}, y_{it}, z_{it})$ to its corresponding spherical coordinate $(r_{it}, \theta_{it}, \delta_{it})$ as follows: $r_{it} = \sqrt{x_{it}^2 + y_{it}^2 + z_{it}^2}$, polar angle $\theta_{it} = \arctan \frac{z_{it}}{\sqrt{x_{it}^2 + y_{it}^2}}$ and azimuthal angle $\delta_{it} = \arctan \frac{y_{it}}{x_{it}}$. Then, we obtain the minimum and maximum values for 3D scanning neighbors' spherical coordinates as follows:

$$\begin{aligned} \theta_{min} &= \min_{p_{it} \in p_{ht}} (\theta_{it}) - 1^\circ, \theta_{max} = \max_{p_{it} \in p_{ht}} (\theta_{it}) + 1^\circ; \\ \delta_{min} &= \min_{p_{it} \in p_{ht}} (\delta_{it}) - 1^\circ, \delta_{max} = \max_{p_{it} \in p_{ht}} (\delta_{it}) + 1^\circ; \end{aligned} \tag{1}$$

Finally, we define the set of 3SN points p_{st} as the subset of points in the environment that satisfy the following conditions:

$$\min_{p_{it} \in p_{ht}} (r_{it}) < r_{st}, \theta_{min} < \theta_{st} < \theta_{max}, \delta_{min} < \delta_{st} < \delta_{max} \tag{2}$$

1° is determined based on the angular resolution (0.2°) of the LiDAR sensor. We adopt the Cartesian coordinates of the 3SN as input.

Overall Pipeline

We establish the LiDAR coordinate system as the global coordinate system, with the origin located at the position of the LiDAR device. For input, we utilize normalized temporal sequence p'_{bt}, p'_{st} and p'_{ht}, which are calculated as follows: $p'_{bt} = p_{bt} - p_{ct}, p'_{st} = p_{st} - p_{ct}, p'_{ht} = p_{ht} - p_{ct}$. The

predicted joint location generated by the entire network is initially in the local coordinate system, and we subsequently transform them into the global coordinate system using p_{ct}.

CoherenceFuse module. To encode p'_{bt}, p'_{st} and p'_{ht}, we employed three separate PointNet++ (Qi et al. 2017) to obtain 3BN features f_b^t, 3SN features f_s^t and human features f_h^t, the size of those three features are all 512 dims. Subsequently, we generated a 1024-dim global frame-wise descriptor f^t by leveraging a fully connected layer to aggregate f_b^t, f_s^t, and f_h^t. To disentangle motion cues and structural edge coherence cues from f_s^t, we employed self-attention layers. This was followed by a concatenation operation, yielding abundant 512-dimensional 3SN features denoted as f_{sa}^t. Further interactions among f_b^t, f_{sa}^t, and f_h^t were modeled using cross-attention layers. After applying a concatenation operation and self-attention layers, we predicted the corresponding joint locations $\hat{\mathbf{J}}^t \in \mathbb{R}^{24\times 3}$.

Kinematics module. Specifically, following ST-GCN (Yan, Xiong, and Lin 2018), we set $\hat{\mathbf{J}}^t$ as a graph node, and the node feature $\mathbf{Q}^t \in \mathbb{R}^{24\times(3+1024)}$ is obtained by concatenating the frame-wise global feature f^t with joint locations $\hat{\mathbf{J}}^t$. The output of ST-GCN is the joint rotations $\mathbf{R}_{6D}^t \in \mathbb{R}^{24\times 6}$. The joint rotations $\mathbf{R}_{6D}^t$ are fed into an off-the-shelf SMPL model to obtain the 24 joints $\hat{\mathbf{J}}_{SMPL}^t \in \mathbb{R}^{24\times 3}$ on the SMPL mesh.

Loss function. 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:

$$\mathcal{L} = \mathcal{L}_{\mathcal{J}} + \mathcal{L}_{\mathbf{\Theta}} + \mathcal{L}_{\mathcal{J}_{SMPL}}, \mathcal{L}_{\mathcal{J}} = \sum_t^T \|\mathbf{J}_{GT}^t - \hat{\mathbf{J}}^t\|_2$$
$$\mathcal{L}_{\mathbf{\Theta}} = \sum_t^T \|\boldsymbol{\theta}_{GT}^t - \hat{\boldsymbol{\theta}}^t\|_2, \mathcal{L}_{\mathcal{J}_{SMPL}} = \sum_t^T \|\mathbf{J}_{GT}^t - \hat{\mathbf{J}}_{SMPL}^t\|_2 \tag{3}$$

where, $\mathbf{J}_{GT}^t$ is the ground truth joint locations of each frame, $\boldsymbol{\theta}_{GT}^t$ is the ground truth pose parameters of the t-th frame.

Experiments

Implementation Details

The training process takes around 100 epochs with Adam optimizer (Kingma and Ba 2015) on one NVIDIA GeForce RTX 3090 Graphics Card. The batch size is set to 8 and the sequence length is set to 16, while the learning rate is set to be 1×10^{-4}. The decay rate is 1×10^{-4}. We set the dropout ratio as 0.1 for the CoherenceFuse module and 0.5 for the ST-GCN module.

Dataset

LidarHuman26M (Li et al. 2022a), **CIMI4D** (Yan et al. 2023) and **SLOPER4D** (Dai et al. 2023) are multi-modal datasets captured using a markless motion capture system, camera, and LiDAR. **LidarHuman26M** records 13 actors performing 20 daily motions. The location of LiDAR is fixed. The collected human data is idealistic without occlusion. We adopt the same dataset split as LidarCap. **CIMI4D** focuses on climbing with heavy self-occlusion. The location of LiDAR is fixed. We randomly split the train and test set based on the sequence. **SLOPER4D** is collected in realistic environments with occlusion and multi-persons standing beside each other. The spatial position of LiDAR changes as the gatherer moves around. We divide each sequence according to 16 frames as a patch, then randomly scramble the patches, and divide the training and testing datasets according to the ratio of 7:3.
Waymo Open Dataset v2.0 (Sun et al. 2019) annotates the location of 14 key points for a single person. It provides 3D point clouds collected by LiDAR and provides RGB images.

Metrics

To evaluate the performance of pose estimation, we report 1) **MPJPE**↓: Mean per root-relative joint position error in millimeters. 2) **PA-MPJPE (PA-)**↓: Aligning predicted skeleton and label with the transformation matrix acquired by Procrusted Analysis, after alignment, calculate MPJPE. 3) **PCK-30 (PCK3)**↑: Percentage of Correct Keypoints with distance to GT lower than 30cm. 4) **PCK-50 (PCK5)**↑: Percentage of Correct Keypoints with distance to GT lower than 50cm. 5) **Accel-error (AE)**↓: Acceleration error between predicted joint and corresponding label point, whose unit is (cm/s^2). 6) **CD**↓: the chamfer distance between the vertices of predicted SMPL mesh and raw point cloud in millimeters.

Comparions

Quantitative. We compare our method to the state-of-the-art learning-based method LidarCap (Li et al. 2022a) which also targets motion capture from LiDAR point clouds and P4T (Fan, Yang, and Kankanhalli 2021) which extracts spatio-temporal features from raw point cloud video to capture motion information. We also compare our method to the optimized-based method HSC2 (Dai et al. 2023) and the image-based method CLIFF (Li et al. 2022b). The results on LidarHuman26M, CIMI4D, and SLOPER4D are shown in Tab.1. For a fair comparison with the HSC2, we set the predicted human motion of the LidarCap as initialized human motion instead of collected human motion from the motion capture system. As described by LidarHuman26M (Li et al. 2022a), due to poor lighting conditions and the low resolution of human bodies, the accuracy of the 2D pose obtained by OpenPose is low. To prevent the misleading optimization direction caused by using an inaccurate 2D pose for projection loss, we removed the projection loss during the training of CLIFF.

On the unoccluded dataset LidarHuman26M, our method improves the MPJPE index by 7.08mm compared with LidarCap. On the dataset CIMI4D with severe self-occlusion, our method improves the MPJPE index by 8.66mm compared with LidarCap. On the more complex dataset SLOPER4D, our method improves the MPJPE index by 5.9mm compared to LidarCap. Since HSC2 uses the distance between the predicted human motion and the raw point clouds, and the physical constraints between humans and scenes to optimize the global location again in the post-optimization stage, our method is inferior to HSC2 in terms of AE index by $8.79cm/s^2$, but it is also better than LidarCap $6.92cm/s^2$.

3BN	3SN	self-attention	CF	MPJPE (mm)↓	PA- (mm)↓	PCK3 (%)↑	PCK5 (%)↑	AE (cm/s^2)↓
×	×	×	×	79.31	66.72	86.00	95.00	45.20
✓	×	×	×	75.30	64.57	86.92	95.35	43.50
×	✓	×	×	75.99	64.75	86.74	95.24	43.37
✓	✓	×	×	74.67	63.33	87.20	95.55	43.84
Stack		×	×	77.87	66.60	86.18	94.98	44.62
✓	✓	✓	×	73.68	62.90	87.40	95.67	39.21
✓	✓	×	✓	**72.23**	**61.67**	**87.94**	**95.79**	**38.28**

Table 2: Ablation study of different components of our framework. CF means CoherenceFuse module. Based on the metrics, shows that each component in the neighborhood-enhanced method is effective.

Radius	MPJPE↓	PA-↓	PCK3↑	PCK5↑	AE↓
2m	**74.47**	**63.34**	**87.31**	**95.71**	**41.59**
3m	76.31	64.45	86.68	95.39	44.04
5m	76.03	64.89	86.69	95.26	43.19
10m	75.24	63.71	87.06	95.50	42.77
20m	75.71	64.67	86.83	95.27	43.05

Table 3: Ablation study of radius parameter in 3BN on Lidarhuman26m. It shows that the two-meter radius parameter is reasonable.

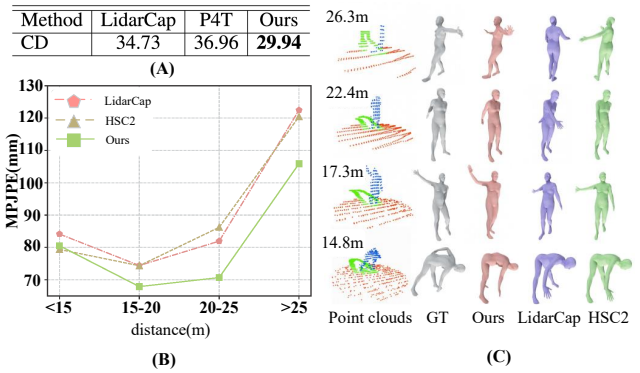

Figure 4: (A): Comparison results on Waymo Open Dataset v2.0 with LidarCap. (B): Comparison with LidarCap, HSC2, and our method in different distances. We show different MPJPE values of different methods. (C): Visualization of methods at different distances.

Generalization. We also compare our method with LidarCap and P4T on the Waymo Open Dataset v2.0. Because Waymo Open Dataset v2.0 does not provide the rotation matrix of each joint which is recorded by the motion capture system, we train our method, LidarCap and P4T on the LidarHuman26M train set and validate it on the Waymo validation set. We extract human points by utilizing the 3D detection box provided by Waymo. In addition, because the number and category of skeleton joints defined by Waymo are inconsistent with SMPL, CD is selected as a quantitative indicator. At the same time, the frame rate of the Waymo Open Dataset v2.0 is 1, while the frame rate of LidarHuman26M is 10. In order to eliminate the influence of different frame rates on the generalization ability, we repeat each frame of Waymo Open Dataset v2.0 16 times as input, and we select the 9th frame of each sequence as output. The results are shown in Fig.4 (A).

On the Waymo Open Dataset v2.0, our method improves the CD index by 4.79mm compared with LidarCap and improves the CD index by 7.02mm compared with P4T, which shows that our algorithm has a stronger generalization ability in real scenes.

Distance analysis. We compare LidarCap and HSC2 with our method in different distances, the results shown in Fig.4 (B). The network achieves significant improvements in both short and long-range motion capture accuracy by simply incorporating 3BN and 3SN. It further highlights the importance of leveraging neighborhood coherence cues in LiDAR-based 3HPE. In the first and last cases, our method estimates better global orientations. We select 4 examples at different distances for visualization which is shown in Fig.4 (C). In the second and third cases, the arm motion predicted by our method is more reliable. Although the estimated human motion slightly differs from the ground truth, our results still outperform others. Based on the metrics and visualization, we conclude that the neighborhood-enhanced method achieves better motion capture results both at close-range and long-range distances.

Qualitative. In Fig.3, we show the qualitative results on LidarHuman26M, CIMI4D, and SLOPER4D with LidarCap and P4T. In the first and second rows, our method estimates reliable motion when compared with the other two methods. Especially the LiDAR point clouds degrade heavily due to distant distance and self-occlusion, the major part of arm and foot point clouds is missing, but our method still manages to estimate more stable human motion. This is achieved by leveraging the structural edge coherence cues in the 3SN. Even though insufficient information is captured on the human surface, the laser rays in contact with the human's edges continue to propagate forward until they collide with the distant environment. As a result, the information about the missing part is recorded in the 3SN. In all cases, our method estimates the joints, which contact with the environment, better than the LidarCap and P4T, we attribute this superiority

to the spatial coherence cues contained in the 3BN.

Ablation Study

Effect of model components. In Tab.2, we ablate the main components of our framework. R1 (Row 1) means baseline method LidarCap (Li et al. 2022a). R2 (Row 2) uses an additional PointNet++ to extract the feature of 3BN and feed it to LidarCap. R3 (Row 3) compared to R2, only replaces the 3BN with the 3SN. Through analysis of the results of R1, R2, and R3, we conclude the 3BN and the 3SN provide more effective cues for the 3HPE network. R4 (Row 4) differs from LidarCap only in the input part, where two independent PointNet++ are added to encode the 3BN and the 3SN, respectively. Then, the features of both are aggregated with the original human feature to obtain global input features. R6 (Row 6) replaces the GRU in LidarCap with a multi-head self-attention layer to fuse these three diverse features, the input is the same as R4, and the whole network is named $LidarCap_s$. R7 (Row 7) replaces the GRU in LidarCap with the CoherenceFuse module.

When comparing R2, R3, R4, and R6, we can see that utilizing a multi-head self-attention layer for information fusion of the three branches achieves the best performance. In R5 (Row 5), compared to LidarCap, there is no difference in the structure of the network. The only difference is that the inputs from the three branches are stacked together and one PointNet++ is used to extract global features. So it is necessary to extract the features of 3BN, 3SN, and individual human points respectively. By comparing R5 and R6, we found the CoherenceFuse module is effective.

Impact of parameters in 3BN. We removed the 3SN branch in the final experimental architecture and experimented by changing the radius range of 3BN on LidarHuman26M. We set 2m, 3m, 5m, 10m, and 20m, a total of 5 radius values, and the results are shown in Tab.3. Because when the radius is set to 1m, the background point clouds that can be extracted are less, so we do not set the experiment with a radius of 1m. By analyzing the data in the table, we can conclude that the information covered in the background has a limited effect on 3HPE. The larger coverage of the background point does not mean that it has more useful information for 3HPE.

Evaluations

To further investigate the information contained in the 3SN which is not merely the outline of human body movements, we conducted a set of comparative experiments with shadows. At first, we obtain plane formulate $ax + by + cz - d = 0$ by fitting a plane in the 3BN. Given a set of points $p_s = \{p_{s1}, p_{s2}, ...p_{sn}\}$ in the 3SN, where $p_{si} = (x_{si}, y_{si}, z_{si})$, we obtain shadow points $p_n = \{p_{n1}, p_{n2}, ...p_{nm}\}$, where $p_{ni} = (k * x_{si}, k * y_{si}, k * z_{si})$ and $k = \frac{d}{(a*x_{si}+b*y_{si}+x*z_{si})}$. In order to investigate the roles played by the 3BN and the 3SN in enhancing the performance of 3HPE, we analyzed the experimental results, as shown in Fig. 5. The curve named 3BN represents 3BN as an additional input to $LidarCap_S$, which is defined in the ablation study. The curve 3BN & 3SN signifies the inclusion of both the 3SN and the 3BN as inputs. The curve 3BN & shadow refers to the inclusion of shadow and 3BN as inputs. The spatial coherence cues in the 3BN contribute to the improvement of the entire scene, while the structural edge coherence cues in the 3SN exhibit a more pronounced enhancement at long distances. Especially in the distant distance, 3SN performs better than shadow claims 3SN is not merely another representation of motion contours. Meanwhile, we discover that the incremental information introduced by 3SN is not as effective as the shadow in close-range scenarios. However, the curve of our method outperforms the other 3 curves. So the CoherenceFuse module enables 3SN to perform well both at close and distant distances.

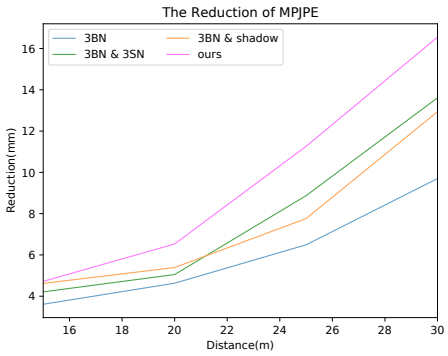

Figure 5: The role of the 3SN and 3BN in different distance scenes. Reduction calculated using the formulation: $\text{MPJPE}_{LidarCap}$ - $\text{MPJPE}_{eachcurve}$.

Conclusion

In this study, we introduced the concept of 3D neighbor in the background (3BN), which leverages spatial coherence cues to enhance the reliability of 3D human pose estimation. Furthermore, we extract 3D scanning neighbors (3SN) based on the inherent properties of LiDAR sensors. These neighbors contribute structural edge coherence cues that facilitate accurate 3D human pose estimation over extended distances, where human LiDAR point clouds are too sparse to supply enough information for reliable estimation. To better integrate the inputs with diverse features, we propose a Transformer-based module named the CoherenceFuse module. Through comprehensive contrast experiments with shadow points, we demonstrate 3SN is not just the outline of motion, but rather a unique characteristic of LiDAR. In addition, our investigations demonstrate that the coherenceFuse module is effective in effectively disentangling the motion cues and structural edge coherence cues within the 3SN. Quantitative and qualitative experiments show that our method outperforms baseline methods, regardless of whether the human subject is at a close or distant distance. However, our current algorithm focuses on processing ideal foreground human point clouds as input, without engaging in comprehensive segmentation or detection tasks across the entire data frame. To address this, our future work will integrate a binary segmentation network into the LiDAR-based 3HPE method, allowing more comprehensive and accurate analysis.

References

Bhatnagar, B. L.; Xie, X.; Petrov, I. A.; Sminchisescu, C.; Theobalt, C.; and Pons-Moll, G. 2022. BEHAVE: Dataset and Method for Tracking Human Object Interactions. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15914–15925.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P. V.; Romero, J.; and Black, M. J. 2016. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *ECCV*.

Cong, P.; Xu, Y.; Ren, Y.; Zhang, J.; Xu, L.; Wang, J.; Yu, J.; and Ma, Y. 2022. Weakly Supervised 3D Multi-person Pose Estimation for Large-scale Scenes based on Monocular Camera and Single LiDAR. *ArXiv*, abs/2211.16951.

Dai, Y.; Lin, Y.; Lin, X.; Wen, C.; Xu, L.; Yi, H.; Shen, S.; Ma, Y.; University, C. W. X.; China; University, S. J. T.; for the Physics of Complex Systems, M. P. I.; and Germany. 2023. SLOPER4D: A Scene-Aware Dataset for Global 4D Human Pose Estimation in Urban Environments. *ArXiv*, abs/2303.09095.

Dai, Y.; Lin, Y.; Wen, C.; Shen, S.; Xu, L.; Yu, J.; Ma, Y.; and Wang, C. 2022. HSC4D: Human-centered 4D Scene Capture in Large-scale Indoor-outdoor Space Using Wearable IMUs and LiDAR.

Fan, H.; Yang, Y.; and Kankanhalli, M. S. 2021. Point 4D Transformer Networks for Spatio-Temporal Modeling in Point Cloud Videos. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14199–14208.

Fürst, M.; Gupta, S. T. P.; Schuster, R.; Wasenmüller, O.; and Stricker, D. 2020. HPERL: 3D Human Pose Estimation from RGB and LiDAR. *2020 25th International Conference on Pattern Recognition (ICPR)*, 7321–7327.

Habermann, M.; Xu, W.; Zollhofer, M.; Pons-Moll, G.; and Theobalt, C. 2020. DeepCap: Monocular Human Performance Capture Using Weak Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hassan, M.; Choutas, V.; Tzionas, D.; and Black, M. J. 2019. Resolving 3D Human Pose Ambiguities With 3D Scene Constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

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

Huang, Y.; Kaufmann, M.; Aksan, E.; Black, M. J.; and Pons-Moll, G. 2018b. Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time.

Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; and Luo, P. 2020. ZoomNAS: Searching for Whole-Body Human Pose Estimation in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 5296–5313.

Joo, H.; Neverova, N.; and Vedaldi, A. 2020. Exemplar Fine-Tuning for 3D Human Model Fitting Towards In-the-Wild 3D Human Pose Estimation. *2021 International Conference on 3D Vision (3DV)*, 42–52.

Kim, H.-W.; Lee, G.-H.; Oh, M.-S.; and Lee, S. 2022. Cross-View Self-fusion for Self-supervised 3D Human Pose Estimation in the Wild. In *Asian Conference on Computer Vision*.

Kim, W.; Ramanagopal, M. S.; Barto, C.; Yu, M.-Y.; Rosaen, K.; Goumas, N.; Vasudevan, R.; and Johnson-Roberson, M. 2019. PedX: Benchmark Dataset for Metric 3-D Pose Estimation of Pedestrians in Complex Urban Intersections. *IEEE Robotics and Automation Letters*, 4: 1940–1947.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, J.; Xu, C.; Chen, Z.; Bian, S.; Yang, L.; and Lu, C. 2020. HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3382–3392.

Li, J.; Zhang, J.; Wang, Z.; Shen, S.; Wen, C.; Ma, Y.; Xu, L.; Yu, J.; and Wang, C. 2022a. LiDARCap: Long-range Markerless 3D Human Motion Capture with LiDAR Point Clouds. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20470–20480.

Li, W.; Liu, H.; Tang, H.; Wang, P.; and Gool, L. V. 2021. MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13137–13146.

Li, Z.; Liu, J.; Zhang, Z.; Xu, S.; and Yan, Y. 2022b. CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation. In *European Conference on Computer Vision*.

Luo, Z.; Iwase, S.; Yuan, Y.; and Kitani, K. 2022. Embodied Scene-aware Human Pose Estimation. In *Advances in Neural Information Processing Systems*.

Marcard, T. V.; Rosenhahn, B.; Black, M. J.; and Pons-Moll, G. 2017. Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs. *Computer Graphics Forum*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, 5099–5108.

Ren, Y.; Zhao, C.; He, Y.; Cong, P.; Liang, H.; Yu, J.; Xu, L.; and Ma, Y. 2022. LiDAR-aid Inertial Poser: Large-scale Human Motion Capture by Sparse Inertial and LiDAR Sensors. *IEEE Transactions on Visualization and Computer Graphics*, 29: 2337–2347.

Robosense. 2023. RS-LiDAR-M1 Brochure EN. https://www.robosense.ai/en/resources. Accessed: 2023-7-14.

Su, Z.; Xu, L.; Zheng, Z.; Yu, T.; Liu, Y.; and Fang, L. 2020. RobustFusion: Human Volumetric Capture with Data-Driven Visual Cues Using a RGBD Camera. In *ECCV*.

Su, Z.; Xu, L.; Zhong, D.; Li, Z.; Deng, F.; Quan, S.; and Fang, L. 2021. RobustFusion: Robust Volumetric Performance Reconstruction Under Human-Object Interactions From Monocular RGBD Stream. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 6196–6213.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; Vasudevan, V.; Han, W.; Ngiam, J.; Zhao, H.; Timofeev, A.; Ettinger, S. M.; Krivokon, M.; Gao, A.; Joshi, A.; Zhang, Y.; Shlens, J.; Chen, Z.; and Anguelov, D. 2019. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2443–2451.

Sun, Y.; Bao, Q.; Liu, W.; Fu, Y.; Black, M. J.; and Mei, T. 2020. Monocular, One-stage, Regression of Multiple 3D People. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11159–11168.

Wan, Z.; Li, Z.; Tian, M.; Liu, J.; Yi, S.; and Li, H. 2021. Encoder-decoder with Multi-level Attention for 3D Human Shape and Pose Estimation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 13013–13022.

Wang, J.; Liu, L.; Xu, W.; Sarkar, K.; Luvizon, D. C.; and Theobalt, C. 2022. Estimating Egocentric 3D Human Pose in the Wild with External Weak Supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13147–13156.

Xie, X.; Bhatnagar, B. L.; and Pons-Moll, G. 2023. Visibility Aware Human-Object Interaction Tracking from Single RGB Camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, L.; Xu, W.; Golyanik, V.; Habermann, M.; Fang, L.; and Theobalt, C. 2020a. 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)*.

Xu, W.; Chatterjee, A.; Zollhöfer, M.; Rhodin, H.; Mehta, D.; Seidel, H.-P.; and Theobalt, C. 2018. MonoPerfCap: Human Performance Capture From Monocular Video. *ACM Transactions on Graphics (TOG)*, 37(2): 27:1–27:15.

Xu, X.; Chen, H.; Moreno-Noguer, F.; Jeni, L. A.; and De la Torre, F. 2020b. 3D Human Shape and Pose from a Single Low-Resolution Image with Self-Supervised Learning. In *ECCV*.

Xu, X.; Chen, H.; Moreno-Noguer, F.; Jeni, L. A.; and De la Torre, F. 2021. 3D Human Pose, Shape and Texture from Low-Resolution Images and Videos. *TPAMI*.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *ArXiv*, abs/2204.12484.

Yan, M.; Wang, X.; Dai, Y.; Shen, S.; Wen, C.; Xu, L.; Ma, Y.; and Wang, C. 2023. CIMI4D: A Large Multimodal Climbing Motion Dataset under Human-scene Interactions. *ArXiv*, abs/2303.17948.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI*.

Ye, D.; Xie, Y.; Chen, W.; Zhou, Z.; and Foroosh, H. 2023. LPFormer: LiDAR Pose Estimation Transformer with Multi-Task Network. *ArXiv*, abs/2306.12525.

Yi, X.; Zhou, Y.; Habermann, M.; Shimada, S.; Golyanik, V.; Theobalt, C.; and Xu, F. 2022. Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors.

Yi, X.; Zhou, Y.; and Xu, F. 2021a. TransPose. *ACM Transactions on Graphics (TOG)*, 40: 1 – 13.

Yi, X.; Zhou, Y.; and Xu, F. 2021b. TransPose: Real-time 3D Human Translation and Pose Estimation with Six Inertial Sensors. *ACM Trans. Graph.*, 40: 86:1–86:13.

Zanfir, A.; Marinoiu, E.; and Sminchisescu, C. 2018. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes: The Importance of Multiple Scene Constraints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2148–2157.

Zhan, Y.-W.; Li, F.; Weng, R.; and Choi, W. 2022. Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13106–13115.

Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; and Yuan, J. 2022a. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13222–13232.

Zhang, J.; Wang, J.; Shi, Y.; Gao, F.; Xu, L.; and Yu, J. 2022b. Mutual Adaptive Reasoning for Monocular 3D Multi-Person Pose Estimation. *Proceedings of the 30th ACM International Conference on Multimedia*.

Zhang, Y.; Wang, C.; Wang, X.; Liu, W.; and Zeng, W. 2021. VoxelTrack: Multi-Person 3D Human Pose Estimation and Tracking in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 2613–2626.

Zheng, J.; Shi, X. Y.; Gorban, A. N.; Mao, J.; Song, Y.; Qi, C.; Liu, T.; Chari, V.; Cornman, A.; Zhou, Y.; Li, C.; and Anguelov, D. 2021. Multi-modal 3D Human Pose Estimation with 2D Weak Supervision in Autonomous Driving. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 4477–4486.