Neurocomputing 712 (2027) 135632

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Reliability-aware quantile-based region aggregation for robust LiDAR-based 3D human pose estimation

Tong Wu[1], Zhongcheng Hong[1], Yanbin Liu, Sira Yongchareon*

School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland, 1142, New Zealand

HIGHLIGHTS

- We propose QuanRA, a reliability-aware framework for robust LiDAR-based 3D HPE.
- Multi-scale Region Structure Learning (MRSL) is developed to suppress noisy points.
- A Quantile-based Dynamic Region Partition (QDRP) handles varying point densities.
- Extensive experiments show state-of-the-art performance on four large-scale benchmarks.

ARTICLE INFO

Communicated by J. Yu

Keywords:
3D human pose estimation
LiDAR
Point cloud
Feature aggregation
Reliability-aware learning

ABSTRACT

LiDAR-based 3D human pose estimation (HPE) remains a difficult problem because point clouds are inherently sparse and noisy. Although recent joint-centric methods aggregate features around joints to handle density variations, they implicitly assume that nearby points are uniformly informative, which often breaks down in cluttered scenes dominated by noisy and irrelevant points. Motivated by this limitation, we revisit the use of point-level features and identify two fundamental challenges: (1) how to assess the reliability of individual points under dominant background clutter, and (2) how to organize reliable point features into structured representations for pose modeling. To address these challenges, we propose the Quantile-based Region Aggregator (QuanRA), a reliability-aware framework for LiDAR-based 3D HPE. To address the first challenge, we propose a Multi-scale Region Structure Learning (MRSL) module for effective point-feature updates. MRSL combines relational features from multi-scale local structures with ego features from self-feature learning to capture semantic differences. A point-level importance estimation mechanism then assigns reliability scores to suppress noisy and irrelevant points. To address the second challenge, we design a Quantile-based Dynamic Region Partition (QDRP) strategy that uses sorted quantile scores to dynamically organize and aggregate reliable points into structured regions under varying point densities. This reliability-aware region formation captures coherent spatial patterns of joint-related evidence. Extensive evaluations on LiDARHuman26M, SLOPER4D, Waymo, and HumanM3 show that our method consistently achieves state-of-the-art performance.

1. Introduction

LiDAR-based 3D human pose estimation (HPE) has attracted increasing attention due to its robustness to illumination changes and its compatibility with autonomous driving and outdoor perception systems [8,43]. Unlike images or depth maps that lie on regular grids, LiDAR data are represented as sparse, noisy point clouds, fundamentally complicating robust 3D HPE [7,18,37,39]. As shown in Fig. 2, in real-world scenarios (such as SLOPER4D [10]), LiDAR point clouds are often noisy, partially occluded, and severely sparse, with human-related points constituting only a small fraction of the entire scene [9,14,26].

Recent advances in 3D HPE have widely adopted joint-centric representation learning [15,44], where joints are treated as semantic aggregation centers to improve pose robustness. Under this paradigm, density-aware methods, such as the density-aware pose transformer

* Corresponding author.
Email addresses: tong.wu@autuni.ac.nz (T. Wu), zhongcheng.hong@autuni.ac.nz (Z. Hong), yanbin.liu@aut.ac.nz (Y. Liu), sira.yongchareon@aut.ac.nz (S. Yongchareon).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.neucom.2026.135632
Received 1 June 2026; Received in revised form 25 September 2026; Accepted 9 November 2026
Available online 10 November 2026
0925-2312/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

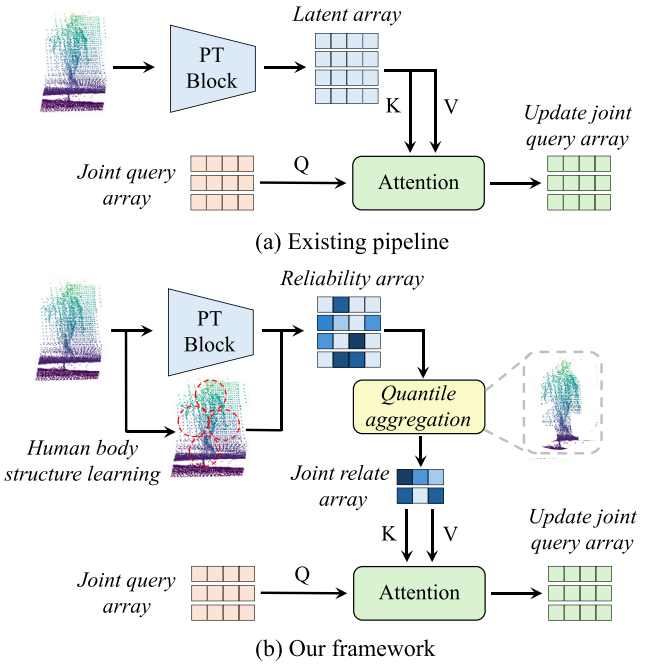


Fig. 1. Real-world LiDAR point clouds are dominated by noisy and irrelevant points. (a) existing joint-centric methods aggregate features from spatially nearby points, implicitly assuming uniform informativeness across local neighborhoods. (b) our framework explicitly estimates *point-level reliability*, organizes reliable evidence into structured regions via *quantile-based aggregation*, and performs joint-level aggregation using only reliable points. It yields more stable joint representations under severe clutter, occlusion, and sparsity.

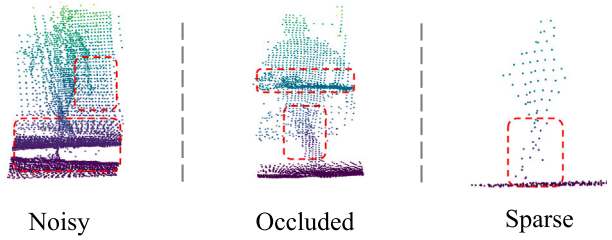


Fig. 2. Challenges in LiDAR-based 3D human pose estimation arise primarily from noisy, occluded, and sparse point cloud samples.

(DAPT) [2] have been introduced for LiDAR-based HPE, which aggregate point features around joints across varying point densities to obtain more stable joint representations. This line of work has demonstrated greater robustness than point-centric regression approaches in sparse point cloud settings.

However, DAPT implicitly assumes that all points in the vicinity of a joint are *equally informative* (cf. Fig. 1a), an assumption that often breaks down in real-world cluttered LiDAR scenes dominated by irrelevant or noisy points. Consequently, joint-level aggregation may become fragile when unreliable point-level features overwhelm the local neighborhood [23,32,38].

Motivated by the above observation, we rethink LiDAR-based 3D HPE from the perspective of point-level reliability and structured modeling. We argue that robust joint-level pose modeling is fundamentally constrained by the reliability of point-level features. In cluttered, noisy environments, it is insufficient to rely solely on a sophisticated joint aggregation mechanism without explicitly assessing which points should contribute to pose estimation. This leads to two fundamental challenges in LiDAR-based 3D HPE: **(1)** how to assess the reliability of individual points under dominant background clutter and **(2)** how to organize reliable point features into structured representations for robust pose modeling.

To address these challenges, we propose the **Quantile-based Region Aggregator (QuanRA)**, a reliability-aware framework (Fig. 1b) for LiDAR-based 3D human pose estimation. QuanRA follows a bottom-up design that progressively refines raw point clouds into structured and reliable representations for pose modeling. *First*, to tackle the challenge of unreliable point-level features, we devise a Multi-scale Region Structure Learning (MRSL) module for effective point feature updates. MRSL captures semantic differences by combining relational features extracted from multi-scale local structures with ego features learned from self-feature representations. Built upon these features, a point-level importance estimation mechanism assigns reliability scores to individual points, enabling the model to suppress noisy and irrelevant background points.

Second, to address the challenge of structured representation learning, we introduce a Quantile-based Dynamic Region Partition (QDRP) strategy. Based on the estimated point-level reliability scores, QDRP uses sorted quantiles to dynamically organize and aggregate reliable points into structured regions under varying point densities. This reliability-aware region formation captures coherent spatial patterns of joint-related evidence, providing stable and informative inputs for downstream pose modeling.

By modeling point-level reliability and reliability-aware region formation, QuanRA enables robust 3D human pose estimation from sparse and noisy LiDAR point clouds.

Our contributions are as follows:

- We propose QuanRA, a reliability-aware framework for LiDAR-based 3D human pose estimation that mitigates the fragility of joint-level aggregation caused by unreliable point-level features.
- We develop a Multi-scale Region Structure Learning (MRSL) module to capture discriminative point-level features, and a Quantile-based Dynamic Region Partition (QDRP) mechanism to construct reliability-aware regions. Together, they improve pose estimation robustness across varying point densities.
- Extensive experiments on LiDARHuman26M, SLOPER4D, Waymo, and HumanM3 show that QuanRA consistently achieves state-of-the-art performance across diverse and challenging scenarios.

2. Related work

2.1. 3D human pose estimation

3D human pose estimation has been extensively studied from monocular images, videos, and multi-view observations. Recent advances have increasingly focused on improving structural reasoning, temporal modeling, geometric consistency, and robustness to ambiguous or occluded observations.

For video-based 3D HPE, Transformer-based approaches have become a dominant paradigm for modeling spatial and temporal dependencies. KTPFormer [27], for example, explicitly incorporates kinematic structure and joint trajectories into attention-based pose modeling, while Hourglass Tokenizer (HoT) [19] improves the efficiency of video pose Transformers by pruning redundant temporal tokens. Other studies further exploit temporal context and scene information to improve pose estimation under severe self-occlusion [1].

Beyond deterministic regression, recent methods have explored generative modeling for handling pose ambiguity. FinePOSE [36] introduces fine-grained part-aware prompts into a diffusion framework for 3D pose estimation, while diffusion-based human recovery methods model a conditional distribution over plausible body configurations [30]. These approaches demonstrate the benefit of richer structural priors when observations are incomplete or ambiguous.

Multi-view 3D HPE has also received considerable attention due to its ability to reduce depth ambiguity through cross-view constraints. MVGFormer [22] combines explicit multi-view geometry with learned

appearance modeling to improve generalization to unseen camera configurations. PoseIRM [3] further addresses robustness under unseen camera settings through invariant representation learning. More recently, MV-SSM [6] introduces state-space modeling to capture joint spatial dependencies across multiple views.

Overall, recent 3D HPE research has substantially improved structural, temporal, and geometric reasoning. However, most of these methods are designed for image- or video-based observations and assume relatively dense visual evidence. LiDAR-based pose estimation presents a different challenge, where human observations are inherently sparse, irregular, and frequently contaminated by background points. These characteristics motivate dedicated representation and aggregation strategies for point-cloud-based 3D HPE.

2.2. LiDAR-based 3D human pose estimation

The practical utility of LiDAR technology has recently sparked a surge in research on point-cloud-based 3D HPE. Pioneering this direction, [18] established a foundational benchmark by releasing the first specialized dataset and a fully supervised baseline, which refines keypoint coordinates via temporal encoding, inverse kinematics, and SMPL models. Subsequent studies have expanded into scene-aware analysis. For instance, [37] introduced a climbing dataset to capture explicit human-scene interactions, while [43] improved pose learning by incorporating environmental background context. In terms of tracking stability, [28,29] demonstrated that exploiting spatiotemporal coherence yields accurate motion results. Alternative approaches focus on multimodal synergy, where researchers [7,13,16,45] leverage 2D image cues and geometric constraints to facilitate weakly supervised training. Furthermore, [40] advanced the field with a multi-task framework, utilizing a keypoint transformer alongside segmentation and detection tasks for multi-person estimation. By learning a stable joint representation, [2] improves the reliability of LiDAR-based 3D human pose estimation while eliminating the need for complex post-processing. Despite notable advances, achieving stable joint representations from sparse and cluttered LiDAR point clouds remains an open challenge. In this work, we explore a reliability-aware perspective that emphasizes robust joint representation learning from single-frame LiDAR data, aiming to improve estimation stability under severe noise and sparsity.

2.3. Point feature aggregation

Feature aggregation is central to point cloud learning for capturing local geometry and global context [41]. Traditional local aggregation methods, such as DeepLA-Net [42], hierarchically merge neighborhood features to improve classification and segmentation. Geometry-guided aggregation further leverages structural cues to propagate local features to global regions, enabling zero-shot recognition [24]. Dynamic and adaptive aggregation strategies have been proposed to overcome the limitations of fixed neighborhoods. For instance, Dynamic Local Feature Aggregation [21] constructs feature-space graphs based on semantic similarity, while Decoupled Local Aggregation [4] separates spatial encoding and aggregation for more efficient local representations. Transformer-based approaches, such as GPSFormer [31] and Point Transformer V3 [35], capture long-range dependencies and integrate local and global information via attention mechanisms. Although recent methods have improved feature aggregation for point clouds, most approaches treat points equally and overlook reliability variations, leading to inconsistent performance under sparse or noisy conditions. Incorporating reliability-aware grouping mechanisms could help dynamically form structured regions of high-confidence points, enhancing robustness and semantic coherence.

3. Methodology

The overall architecture of the LiDAR-based pose estimation framework is shown in Fig. 3. Specifically, we propose a novel Quantile-based Region Aggregator (QuanRA) to progressively suppress noisy and irrelevant points while sampling and aggregating key points to represent

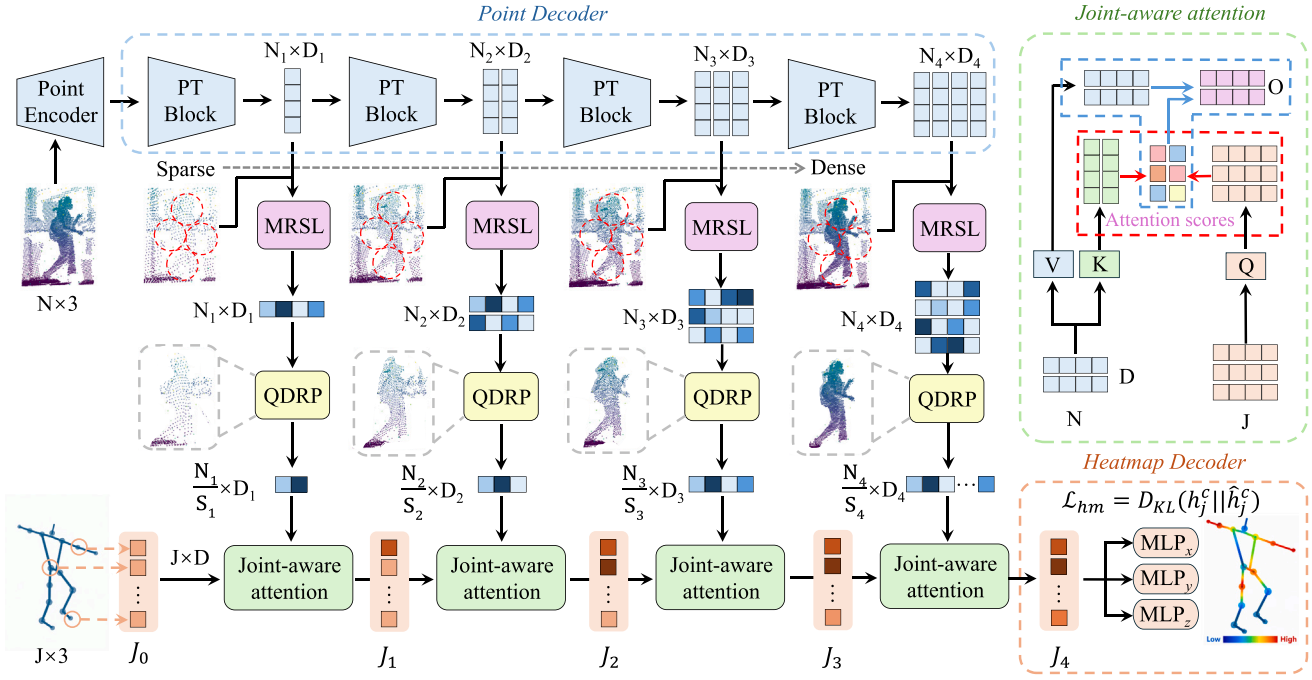


Fig. 3. Illustration of our Quantile-based Region Aggregator (**QuanRA**) for LiDAR-based 3D human pose estimation. QuanRA adopts a *reliability-aware* design: point-level importance is first estimated via Multi-scale Region Structure Learning (**MRSL**); then reliable points are dynamically organized into structured regions using Quantile-based Dynamic Region Partition (**QDRP**). Joint-level feature aggregation is subsequently performed through joint-aware attention, and the resulting joint representations are decoded into coordinate-wise heatmaps by a lightweight heatmap decoder. This design enables *stable and robust pose estimation* under severe noise, occlusion, and point sparsity.

the human body structure. The decoder utilizes these key points to efficiently update the keypoint features, which serve as queries for human skeleton pose. QuanRA consists of a cascade of modules, including the Point Neighbor Graph, Multi-scale Region Structure Learning (MRSL), and Quantile-based Dynamic Region Partition (QDRP), which are described in detail in the following sections.

3.1. Quantile-based region aggregator (QuanRA)

Given a noisy point cloud $\mathcal{P} = \{p_i\}_{i=1}^{N}$ with N points, where $p_i = (x, y, z) \in \mathbb{R}^3$ denotes the 3D coordinates. Similar to DAPT [2], we first encode the point cloud into sparse point features using a point encoder, and then feed these features into a decoder to reconstruct point-level features. During decoding, we design an innovative quantile-region aggregator to select and aggregate informative points from noisy point clouds, which are then used for precise pose estimation.

3.1.1. Point-neighbor graph

QuanRA first constructs a point-neighbor graph to capture local geometric variations in pose. Specifically, each embedded feature produced by a point transformer block is first converted into a point representation for graph construction. Then, for each point p_i, we construct a k-nearest neighbor graph whose nodes are $\{p_i^j\}_{j=1}^{k}$. Then, we compute the node features as feature differences *w.r.t.* p_i: $\mathcal{K}(p_i) = \{f_{p_i} - f_{p_i}^j | j = 1, \dots, k\}$. Here, f_{p_i} denotes the coordinates of p_i for the first block, while for later blocks b, we use the features from the previous block $b-1$. The whole graph can be represented as

$$f_{\mathcal{P}} = \{(f_{p_i}, \mathcal{K}(p_i))\}_{i=1}^{N}, \tag{1}$$

where $f_{\mathcal{P}} \in \mathbb{R}^{N \times D \times k}$ represents the point-neighbor graph features with dimension D. We adopt *feature differences* instead of the original features because the relative movement, rather than the absolute position, matters in pose estimation.

3.1.2. Multi-scale region structure learning (MRSL)

With the neighbor graph, we then design the MRSL module to perform feature updates. MRSL considers a point-neighbor graph $f_{\mathcal{P}}$ to capture subtle feature variations and contextual information within a local region.

As shown in Fig. 4, we apply multi-scale neighbor feature aggregation (MNFA) to augment point neighbor graph features $f_{\mathcal{P}}$, which enables the model to accurately distinguish different local structures under varying noise and point cloud densities. Then, we use a multi-layer perceptron (MLP) and max pooling to downscale and encode multi-scale local structural information to obtain multi-scale graph features $f_{\mathcal{MP}}$. MNFA takes into account neighboring information, but this might weaken the ego representation of the specific center point, especially when it is a body joint. Therefore, we propose self-feature learning (SFL) to counteract this effect and obtain ego features $f_{\mathcal{EG}}$. Finally, we perform element-wise summation of these two features and adaptive fusion via attention feature integration (AFI) to further improve the model's overall feature representation. The whole process can be represented as

$$f_{\mathcal{MP}} = \text{MaxPooling}(\text{MLP}_1(\text{MNFA}(f_{\mathcal{P}}))), \tag{2}$$

$$f_{\mathcal{EG}} = \text{MLP}_2(\text{SFL}(f_{\mathcal{P}})), \tag{3}$$

$$\mathcal{V}_{\mathcal{P}} = \sigma(\text{DS-Conv}(f_{\mathcal{MP}} + f_{\mathcal{EG}})), \tag{4}$$

$$f'_{\mathcal{P}} = \mathcal{V}_{\mathcal{P}} \odot \text{SA}(f_{\mathcal{MP}}) + (1 - \mathcal{V}_{\mathcal{P}}) \odot \text{CA}(f_{\mathcal{EG}}). \tag{5}$$

Here, MLP_1 and MLP_2 denote two independently parameterized MLPs with non-shared weights. MLP_1 transforms the multi-scale neighborhood features produced by MNFA before max pooling, thereby encoding relational information between each center point and its local neighborhood. In contrast, MLP_2 operates on the SFL branch to encode the ego representation of the center point. The two branches therefore capture complementary local-relational and point-specific information, respectively, before being adaptively integrated by AFI. DS-Conv stands for Deep Separable Convolution, σ represents the sigmoid function, and $\mathcal{V}$ represents the compressed attention vectors of $f_{\mathcal{MP}}$ and $f_{\mathcal{EG}}$ after element-wise summation. SA$(\cdot)$ and CA$(\cdot)$ represent spatial and channel attention modules, respectively, which filter unimportant information in the features and assign higher weights to important center points and their corresponding local regions. $f'_{\mathcal{P}} \in \mathbb{R}^{N \times D}$ denotes the final feature representation.

3.1.3. Quantile-based dynamic region partition (QDRP)

Given the feature representation $f'_{\mathcal{P}}$ of all points in a cloud, a key problem is how to partition and sample points in an efficient and effective manner. For example, MM-Fi [39] adopts a uniform strategy by uniformly partitioning and sampling. Top-S [5] selects the top S points as centers for partitioning and sampling. These strategies are not specifically designed for human pose estimation, and thus achieve suboptimal performance.

We propose a Quantile-based Dynamic Region Partition (QDRP) strategy (Fig. 5) tailored for pose estimation which captures the distributional changes of dynamic points in a human point cloud.

Here, point-level reliability refers to the estimated contribution of each point to human pose estimation. In QDRP, this reliability is represented by an importance score predicted from the point feature: a larger score indicates that the corresponding point provides more informative evidence for pose estimation.

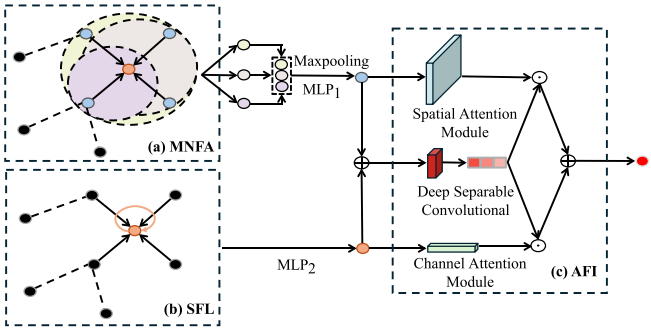


Fig. 4. Multi-scale Region Structure Learning (MRSL) integrates multi-scale local structure (MNFA) with ego features (SFL) and fuses them through an attention feature integration (AFI) module to refine point-level representations.

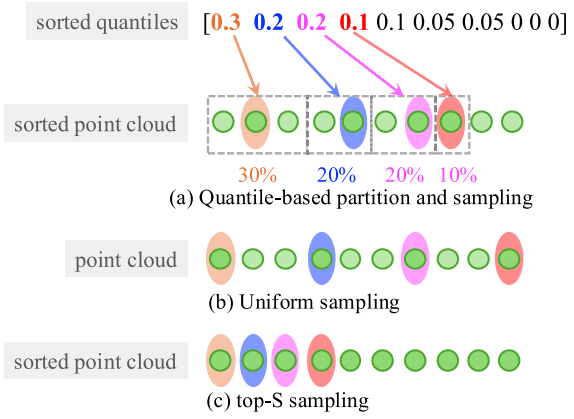


Fig. 5. (a) Our Quantile-based Dynamic Region partition (QDRP) first partitions the points according to sorted quantiles, then samples points within each partition. (b) and (c) Other methods either use a uniform partition strategy [39] or a top-S partition strategy [5].

First, we calculate the reliability (importance) scores for all points using the updated features f'_p in MRSL:

$$\mathcal{A},\ \mathcal{I} = \mathrm{Sort}\left(\sigma\left(\mathrm{MLP}\left(f'_{\mathcal{P}}\right)\right)\right), \tag{6}$$

where $\mathcal{A}$ denotes the sorted scores, and $\mathcal{I}$ denotes the sorted indices.

Specifically, $\mathcal{A}_i$ is the point-level reliability score of the i-th point after sorting, and the points are arranged in descending order of reliability. Thus, the importance score and the point-level reliability refer to the same quantity in our formulation.

Then, we normalize the scores to obtain the reliability distribution over all sorted points as:

$$\mathcal{Q}_i = \frac{\mathcal{A}_i}{\sum_{j=1}^{N} \mathcal{A}_j}, \quad i \in \{1, 2, \ldots, N\}, \tag{7}$$

where $\mathcal{Q}_i$ denotes the normalized reliability weight of each sorted point, satisfying $\sum_{i=1}^{N} \mathcal{Q}_i = 1$.

To determine the quantile boundaries, we further construct the cumulative reliability distribution:

$$\mathcal{C}_i = \sum_{j=1}^{i} \mathcal{Q}_j, \quad i \in \{1, 2, \ldots, N\}, \tag{8}$$

where $\mathcal{C}_i$ denotes the cumulative reliability quantile up to the i-th sorted point. Given S partition regions, the boundary of the s-th region is determined by

$$b_s = \min\left\{ i \mid \mathcal{C}_i \geq \frac{s}{S} \right\}, \quad s \in \{1, 2, \ldots, S\}, \tag{9}$$

with $b_0 = 0$ and $b_S = N$. Accordingly, the s-th region consists of the sorted points whose indices lie between $b_{s-1}+1$ and b_s. Therefore, the region boundaries are determined by cumulative reliability rather than by a fixed number of points, allowing the partition to adapt to the reliability distribution of each input point cloud.

In this way, the predicted importance scores first define point-level reliability, the normalized scores form a reliability distribution, and the cumulative quantiles determine the dynamic partition boundaries.

The center point is randomly selected in each region and then aggregated.

In Fig. 5, we show a comparison of sampling four points out of ten. After completing the regional segmentation, we sample representative points from the delineated regions to avoid computationally intensive dense graph processing.

Specifically, sampling is performed within each reliability-aware region defined by the quantile boundaries above.

These representative points are then used as the Key (K) and Value (V), respectively, which jointly attend to the Query (Q) human skeleton features to update the features. The whole process can be represented as

$$\mathcal{X}_s = \texttt{Sample}\left(\mathcal{P}_{I[b_{s-1}+1:b_s]}\right), \quad s \in \{1, \ldots, S\}, \tag{10}$$

$$f'_{\mathcal{X}_s} = f'_{\mathcal{P}_{(\mathcal{X}_s)}} \odot \mathcal{A}_{(\mathcal{X}_s)}, \quad s \in \{1, \ldots, S\}, \tag{11}$$

$$f'_{\mathcal{J}} = \texttt{Joint-aware_attention}\left(\left\{f'_{\mathcal{X}_s}\right\}_{s=1}^{S}, f_{\mathcal{J}}\right). \tag{12}$$

where S denotes the number of partition regions, $\mathcal{X}_s$ represents the set of points sampled from the s-th reliability-aware region, and $\mathcal{A}_{(\mathcal{X}_s)}$ denotes their corresponding reliability scores. $f'_{\mathcal{X}_s}$ denotes weighted point features, J represents the number of human joints, and $f_{\mathcal{J}}, f'_{\mathcal{J}} \in \mathbb{R}^{J\times D}$ denote the joint features before and after joint-aware attention, respectively.

3.2. Heatmap decoder

As shown in Fig. 3, the joint-aware attention outputs the final human skeleton features, which are decoded into 3D joint coordinates via an axis-wise heatmap formulation [20]. For each joint, three lightweight MLP heads predict 1D heatmaps along the x, y, and z axes, respectively. Each heatmap represents the probability distribution of the joint location within a predefined coordinate range. The final joint coordinates are obtained by decoding the predicted heatmap peaks and mapping them back to continuous space.

To adapt the model to dataset-specific joint annotations, we supervise the predicted heatmaps using the Kullback-Leibler divergence. The heatmap loss is defined as

$$\mathcal{L}_{\mathrm{hm}} = \sum_{j=1}^{K+1} \sum_{c\in\{x,y,z\}} D_{\mathrm{KL}}\left(h_j^c \,\|\, \hat{h}_j^c\right), \tag{13}$$

where h_j^c and $\hat{h}_j^c$ denote the ground-truth and predicted 1D heatmaps of the j-th joint along axis c, respectively.

4. Experiment

4.1. Implementation details

For model configuration, we adopt the standard PTv3 setup [35], using a voxelization grid size of 0.01. Similar to DAPT [2], we first pre-train the point encoder and decoder using synthetic data. The pre-training process follows the official implementation, with no additional heuristics introduced.

For a fair comparison, LPFormer, PRN, DAPT, and all QuanRA variants use an identical backbone and training protocol. All models are optimized using AdamW with a weight decay of 0.005 and a OneCycleLR schedule with cosine annealing, and are fine-tuned for 50 epochs. The main experimental results are obtained using a fixed random seed of 42. For the additional experiments introduced in response to the reviewers, we repeat the experiments using three random seeds, $\{42, 43, 44\}$, and report the corresponding mean and standard deviation where applicable.

Following the synthetic pre-training strategy of DAPT [2], all models are first pre-trained on synthetic LiDAR scans generated from SMPL human bodies before dataset-specific fine-tuning. Synthetic scans are generated by ray casting against the SMPL mesh with randomized sensor orientations and distances ranging from 5 to 20 m. We use a ray keep rate of 0.6 and apply two occlusion augmentations, including clustered noise-point injection and random removal of one SMPL body-part segment, each with a probability of 0.5.

The dataset-specific training configurations are kept fixed across all compared methods. LiDARHuman26M uses a batch size of 32, with temporal subsampling intervals of 5 and 1 for training and testing, respectively. Waymo Open Dataset v2 uses a batch size of 64 without temporal subsampling. HumanM3 uses a batch size of 64 with a temporal subsampling interval of 10, while SLOPER4D uses a batch size of 64 without additional temporal subsampling.

The number of neighbors k is set to 20, and the region-partition configuration S is set to $[2, 4, 8, 16]$ across the four decoder stages, respectively.

4.2. Datasets

We evaluate our method on four public datasets that differ in sensing setups, scene complexity, and annotation protocols.

LiDARHuman26M [18] is a multi-modal dataset where human poses are captured using inertial measurement units (IMUs) and provided in SMPL format. A fixed LiDAR sensor records point clouds of various daily activities within the distance range of 12–24 meters. The dataset features clean and controlled scenes, making it suitable for evaluating baseline performance.

We use the official training and test split provided by the dataset. During training, frames are temporally subsampled at an interval of 5, while the test set is evaluated at an interval of 1.

SLOPER4D [10] provides IMU-annotated human poses captured in more realistic environments. It employs a mobile LiDAR sensor to follow a walking subject and collect point cloud data.

Table 1
Comparison with other methods on the IMU-annotated LiDARHuman26M, SLOPER4D, and the manually annotated Waymo Open Dataset and HumanM3. PA- denotes PA-MPJPE, and the units of MPJPE and PA-MPJPE are millimeters (mm). Best results are highlighted in bold. For each dataset, the skeleton structures are aligned to ensure fair comparison.

Dataset	Waymo			HumanM3			LiDARHuman26M				SLOPER4D			
Metrics	MPJPE↓	PCK3↑	PCK5↑	MPJPE↓	PCK3↑	PCK5↑	MPJPE↓	PA-↓	PCK3↑	PCK5↑	MPJPE↓	PA-↓	PCK3↑	PCK5↑
LidarCap	–	–	–	–	–	–	79.31	66.72	86.00	95.00	101.89	78.93	78.15	89.77
NE	–	–	–	–	–	–	72.23	61.67	87.94	95.79	96.80	76.70	79.22	90.51
LPFormer	61.60	94.52	98.04	83.69	89.54	96.85	95.72	79.03	84.38	94.87	49.31	38.92	97.22	99.44
PRN	68.48	93.60	97.87	70.56	93.18	97.80	80.01	63.67	88.77	96.54	48.76	39.42	97.38	99.45
DAPT	51.59	97.34	98.98	59.76	95.39	98.30	73.47	58.40	90.50	97.06	28.01	21.52	99.30	99.87
QuanRA (ours)	**48.13**	**98.04**	**99.19**	**55.64**	**96.62**	**99.02**	**68.86**	**54.44**	**91.78**	**97.58**	**26.15**	**19.84**	**99.37**	**99.89**

Since SLOPER4D does not provide an official train–test split, we follow the partitioning protocol of Zhang et al. [43]. Specifically, each sequence is divided into 16-frame patches, from which a fixed 70/30 train–test split is generated once and cached. The same split is used for all baselines and QuanRA variants to ensure a controlled comparison. No additional temporal subsampling is applied.

HumanM3 [11] is a large-scale multi-person pose dataset with annotations generated through automatic labeling followed by manual refinement. It mainly consists of sports-field scenarios, where point clouds and images are captured by multiple synchronized RGB–LiDAR sensor pairs.

For reproducibility, we use the official `train.pkl` and `test.pkl` files generated by the HumanM3 extraction pipeline. Due to the high temporal redundancy of adjacent frames, the training data are temporally subsampled with an interval of 10.

Waymo Open Dataset v2 [25] is a large-scale autonomous driving dataset that includes manually annotated 3D human poses. It contains over 10 K human instances collected in diverse real-world outdoor environments.

We use the official training and validation sequence lists for model training and evaluation, respectively, without additional temporal subsampling.

4.3. Quantitative results

We compare our method with several state-of-the-art approaches. Specifically, we evaluate recent Transformer-based models, including LPFormer [40], PRN [12], and DAPT [2], all of which infer full-body pose from a single LiDAR frame. We also include LiDARCap [18], which leverages SMPL optimization and temporal information, as well as the neighborhood-enhanced NE-3HPE [43] that incorporates surrounding scene context. Table 1 summarizes the quantitative results on four benchmarks, with a particular focus on comparisons against DAPT. Our method consistently outperforms DAPT across all datasets and evaluation metrics. On Waymo, our approach reduces MPJPE from 51.59 mm to 48.13 mm, while simultaneously improving PCK3 and PCK5, demonstrating more accurate pose estimation in autonomous driving scenarios. On HumanM3, which emphasizes complex human motions, our method achieves an MPJPE of 55.64 mm, outperforming DAPT by 4.12 mm. On LiDARHuman26M, our approach yields substantial gains over DAPT, reducing MPJPE from 73.47 mm to 68.86 mm and PA-MPJPE from 58.40 mm to 54.44 mm, indicating improved global alignment. Most notably, on the challenging SLOPER4D benchmark, our method dramatically lowers MPJPE from 28.01 mm to 26.15 mm. Overall, our method consistently achieves higher PCK scores than those of DAPT across all datasets, validating its robustness and effectiveness.

4.4. Error distribution analysis

To further investigate the effectiveness of the proposed QuanRA framework, we conduct a detailed analysis of the per-joint error distribution on the SLOPER4D dataset, as illustrated in Fig. 7. The visualization presents the spatial distribution of the Mean Per Joint Position Error (MPJPE) across different body joints, where cooler colors indicate higher localization errors and warmer colors denote more accurate predictions.

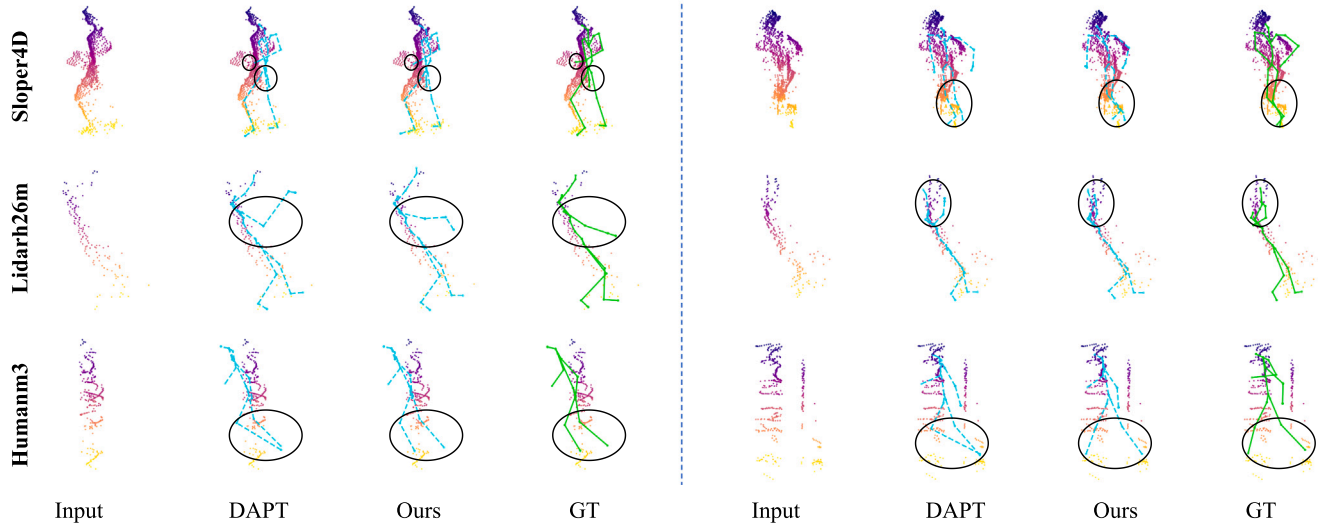


Fig. 6. Visualization results on standard datasets with challenging samples show that our method maintains strong robustness, even on low-quality point clouds affected by occlusion, noise, and sparsity.

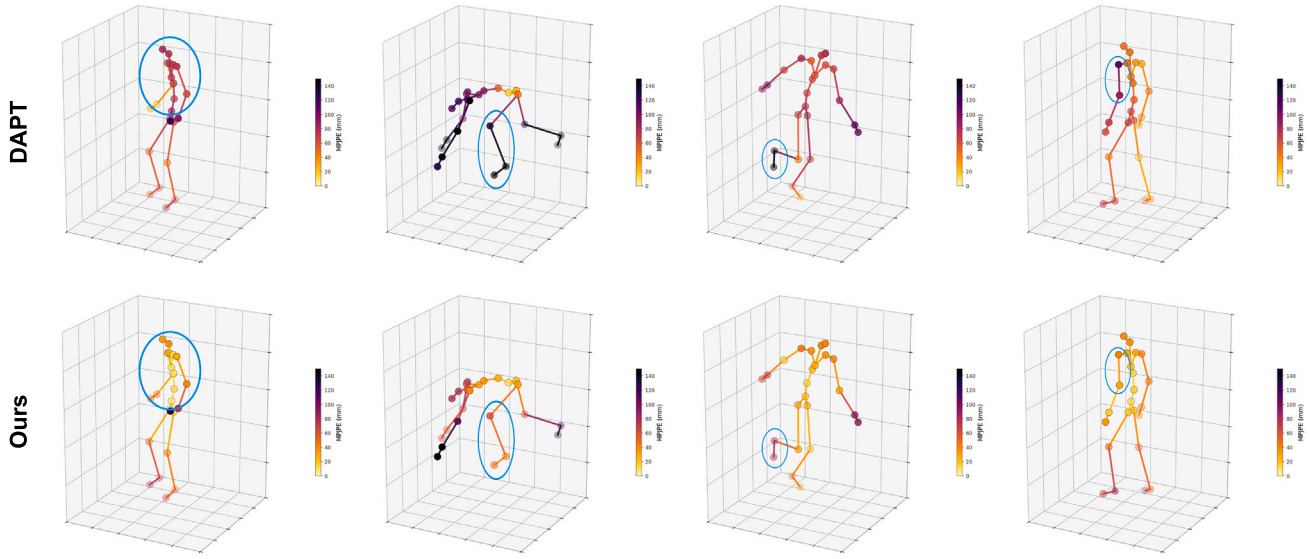


Fig. 7. Error distribution heatmap analysis on the SLOPER4D datasets. The heatmap colors indicate the per-joint MPJPE (Mean Per-Joint Position Error), where cooler colors (blue) represent higher errors and warmer colors (red) indicate more precise estimations. Compared with the baseline DAPT, our method significantly reduces the localization errors in key anatomical joints, particularly the shoulders, knees, and feet, which are often prone to structural ambiguity in sparse point clouds.

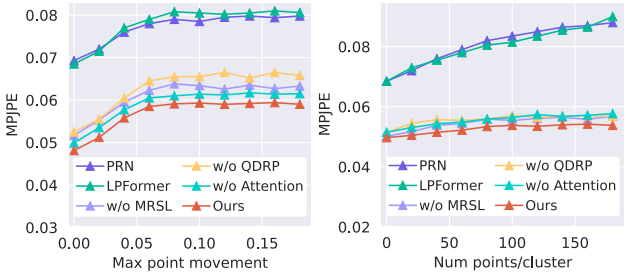


Fig. 8. Stability evaluation of our method. The proposed approach exhibits limited performance degradation under perturbations, remaining within an acceptable range.

Compared with the baseline DAPT, our method consistently exhibits reduced errors across most joints. Notably, significant improvements are observed in anatomically challenging regions, including the shoulders, knees, and feet, which are particularly sensitive to noise and structural ambiguity in sparse LiDAR point clouds.

As shown in Fig. 7, DAPT tends to produce larger errors in these regions, suggesting that its joint-centric aggregation strategy is vulnerable to unreliable or noisy points in local neighborhoods. In contrast, our method demonstrates more stable and accurate joint localization. This is attributed to QuanRA's reliability-aware design, which explicitly models point-level importance and mitigates the influence of irrelevant observations.

Specifically, the Multi-scale Region Structure Learning (MRSL) module enhances feature representation by capturing discriminative local structures at multiple scales, while the Quantile-based Dynamic Region Partition (QDRP) adaptively organizes points according to their reliability. This combination enables the model to emphasize structurally meaningful regions and suppress noisy or misleading points, leading to improved robustness in challenging scenarios. Overall, the results in Fig. 7 indicate that the proposed method not only improves overall accuracy but also achieves more balanced error distributions across joints. This is particularly important for practical applications, where reliable estimation of critical joints is essential for downstream tasks such as motion analysis and human behavior understanding.

4.5. Stability evaluation

To evaluate the stability of our method, we introduce controlled disturbances to the input point cloud and re-evaluate model performance. Specifically, we consider point jittering and noise point clusters and report the results in Fig. 8. To provide a more detailed quantitative evaluation, we further report the robustness results across six perturbation levels for both point jittering and clustered-noise corruption in Table 4. For each perturbation setting, the experiment is repeated three times, and the results are reported as the mean ± standard deviation. QuanRA consistently achieves lower MPJPE than the baseline under all tested perturbation levels. In particular, under point jittering, the relative improvement reaches 15.5% at the strongest perturbation level, while under clustered-noise perturbations QuanRA also maintains consistently better performance.

Point jittering. We perturb each input point by adding random positional offsets with increasing maximum magnitudes. As shown in

Table 2
Ablation study with progressively enabled components.

MRSL	QDRP	Joint-aware Attention	MPJPE↓
DAPT Baseline			73.4
✓			71.8 (−1.6)
✓	✓		69.7 (−3.7)
✓		✓	71.3 (−2.1)
	✓	✓	70.5 (−2.9)
✓	✓	✓	**68.9** (−4.5)

Table 3
Ablation study of MRSL and AFI modules. w/means with (having); w/o means without (not having).

Scale Strategy	Attention (AFI)	MPJPE (mm)↓
Single-scale	w/o	22.50
Multi-scale (MRSL)	w/o	20.60 (−1.90)
Multi-scale (MRSL)	w/	**19.84** (−2.66)

Table 4
Quantitative robustness evaluation under different levels of point jittering and clustered-noise perturbations on SLOPER4D. Results are reported as the mean ± standard deviation over three seeds (42, 43, and 44).

Point Jittering				Noise Point Clusters			
Magnitude	Baseline MPJPE ↓	QuanRA MPJPE ↓	Relative Gain	Points/Cluster	Baseline MPJPE ↓	QuanRA MPJPE ↓	Relative Gain
0.00	28.01 ± 0.56	**26.15 ± 0.52**	6.6%	0	28.01 ± 0.56	**26.15 ± 0.52**	6.6%
0.01	27.60 ± 0.55	**25.85 ± 0.53**	6.7%	20	37.83 ± 1.33	**37.62 ± 1.13**	0.5%
0.02	28.55 ± 0.71	**26.40 ± 0.66**	7.5%	50	45.43 ± 1.59	**44.94 ± 1.57**	1.1%
0.05	46.93 ± 1.64	**42.46 ± 1.49**	9.5%	100	51.59 ± 2.06	**48.18 ± 1.93**	6.6%
0.08	81.01 ± 4.05	**76.94 ± 3.85**	5.0%	150	58.08 ± 2.61	**52.39 ± 2.36**	9.8%
0.12	137.55 ± 8.25	**116.23 ± 6.97**	15.5%	200	60.95 ± 3.05	**55.73 ± 2.79**	8.6%

Fig. 8(a), our method exhibits noticeably smaller performance degradation compared with that of the baseline approaches. In particular, removing MRSL leads to a more pronounced increase in MPJPE under strong jitter, indicating that MRSL effectively enforces local structural consistency and mitigates the impact of positional perturbations.

Noise point clusters. We further evaluate robustness by injecting clusters containing varying numbers of noise points into the input point cloud. As illustrated in Fig. 8(b), the full model maintains stable performance as the number of noise points per cluster increases. In contrast, the variant without QDRP suffers from more severe degradation, suggesting that QDRP plays a critical role in handling clustered noise by suppressing low-quality or noisy point distributions.

4.6. Ablation study

Model components. Table 2 presents the ablation study by progressively enabling each component on LiDARHuman26M. Starting from the DAPT baseline, introducing MRSL leads to a clear performance gain, reducing MPJPE from 73.4 mm to 71.8 mm, indicating the effectiveness of enforcing structural consistency. Incorporating QDRP further improves the results, achieving an MPJPE of 69.7 mm, which suggests that reliability-aware distribution modeling complements MRSL. Adding joint-aware attention yields additional performance gains, reducing MPJPE to 70.5 mm when combined with QDRP and to 71.3 mm when combined with MRSL. Finally, enabling all components together achieves the best performance with an MPJPE of 68.9 mm, demonstrating the complementary nature of MRSL, QDRP, and joint-aware attention.

We also conduct an ablation study to verify the effectiveness of the Multi-scale Region Structure Learning (MRSL) and the Attention Feature Interaction (AFI) module. As shown in Table 3, replacing the multi-scale approach with a single-scale baseline results in a performance drop of 1.9 mm. Furthermore, removing the AFI (Attention Feature Interaction) module degrades precision, highlighting the importance of feature refinement.

Neighbor search strategy. Table 5 compares different neighbor search strategies on SLOPER4D. The experimental results demonstrate that KNN consistently outperforms radius-based search, likely because KNN provides a stable number of neighboring points regardless of the non-uniform density often found in human point clouds, whereas radius search may capture an inconsistent amount of local information. Among the KNN configurations, $k = 20$ achieves optimal performance with a PA-MPJPE of 19.84 mm, suggesting it strikes the ideal balance between capturing sufficient local geometric context and avoiding noise from irrelevant body parts or the background. Specifically, a smaller neighborhood ($k = 16$) lacks the necessary spatial information for accurate pose estimation, while a larger neighborhood ($k = 24$) introduces excessive noise that degrades the precision of the local representation.

Overall, these findings indicate that while our framework maintains relatively stable performance across reasonable variations of k, selecting an appropriate fixed-count neighbor strategy is essential for robust feature extraction in 3D human pose estimation.

Region partition setting. Table 6 compares different region partition settings on HumanM3. We find that the partition $[2, 4, 8, 16]$ achieves the highest PCK3 and PCK5, indicating that progressively increasing the region resolution effectively captures both coarse and fine geometric details. This superiority suggests that gradually increasing the region resolution allows the model to effectively capture both coarse global structures and fine-grained local geometric details. Conversely, the regressive setting performs significantly worse, underscoring the importance of maintaining higher resolution in the deeper network stages to refine point-level representations. These results also suggest that our method is robust to reasonable variations in partition design.

Sampling strategy. To isolate the contribution of QDRP, we replace its quantile-based partitioning with random, importance-weighted, and top-k sampling while keeping all other components unchanged. As shown in Table 8, QDRP achieves the lowest MPJPE of 26.15 ± 0.52 mm, compared with 31.64 ± 1.11 mm for random sampling, 38.12 ± 1.52 mm for importance-weighted sampling, and 39.29 ± 1.57 mm for top-k sampling. This demonstrates that quantile-based region partitioning is more effective than direct importance-based point selection.

MRSL components. We further conduct a fine-grained ablation of MNFA, SFL, and AFI. As shown in Table 9, removing MNFA, SFL, and AFI increases MPJPE from 26.15 mm to 26.86 mm, 26.32 mm, and 27.73 mm, respectively. These results show that all three components contribute to MRSL, while AFI provides the largest individual improvement by adaptively integrating the multi-scale neighborhood and self-feature representations.

Comparison with adaptive sampling methods. We further compare QDRP with PointSP [17], APES [34], and SAMBLE [33] by replacing only the sampling strategy while keeping MRSL and the remaining architecture unchanged. As shown in Table 10, QDRP achieves the best overall performance, with 26.15 mm MPJPE and 19.84 mm PA-MPJPE. These results demonstrate the effectiveness of reliability-aware quantile-based region construction over existing adaptive point-sampling strategies.

Table 5
Comparison of different neighbor search strategies.

Method	PA-MPJPE↓
KNN (k = 16)	23.57
KNN (k = 20)	**19.84**
KNN (k = 24)	21.82
Radius Neighbors (r = 0.1)	27.54
Radius Neighbors (r = 0.2)	25.68

Table 6
Comparison of different region partition settings.

Partition Ration S	PCK3↑	PCK5↑
[4, 4, 4, 4]	95.41	98.36
[16, 8, 4, 2]	93.84	97.78
[1, 2, 4, 8]	95.76	98.59
[2, 4, 8, 16]	**96.62**	**99.02**

Table 7
Efficiency–performance comparison of different sampling strategies on 270 real SLOPER4D test frames with a batch size of 1. FLOPs and latency are reported as the mean ± standard deviation.

Method	Params	GFLOPs/frame	Peak GPU memory	Latency/frame	MPJPE ↓
QDRP (ours)	64.7580 M	34.535 ± 6.792	357.7 MB	22.76 ± 1.71 ms	**26.15**
PointSP	64.7571 M	34.535 ± 6.792	357.7 MB	21.67 ± 1.61 ms	27.07
APES	65.4452 M	34.631 ± 6.814	360.4 MB	21.00 ± 1.48 ms	34.99
SAMBLE	65.4488 M	34.579 ± 6.800	360.4 MB	20.75 ± 1.41 ms	36.33

Table 8
Comparison of different sampling strategies. Results are reported as mean ± standard deviation over three seeds (42, 43, and 44).

Sampling strategy	MPJPE ↓
QDRP (quantile)	**26.15 ± 0.52**
Random	31.64 ± 1.11
Importance-weighted	38.12 ± 1.52
Top-k	39.29 ± 1.57

Table 9
Fine-grained ablation of MRSL components on SLOPER4D.

Variant	MPJPE ↓
Full model	**26.15**
w/o MNFA	26.86
w/o SFL	26.32
w/o AFI	27.73
Baseline (w/o MRSL and QDRP)	28.01

Table 10
Comparison with adaptive point-sampling methods on SLOPER4D.

Method	MPJPE ↓	PA-MPJPE ↓	PCK3 ↑	PCK5 ↑
Baseline (DAPT)	28.01	21.52	99.30	99.87
PointSP	27.07	21.74	98.59	99.73
APES	34.99	22.89	97.77	99.81
SAMBLE	36.33	21.84	98.03	99.76
QDRP (ours)	**26.15**	**19.84**	**99.37**	**99.89**

4.7. Visualization

Prediction results. Fig. 6 showcases the pose estimation results on three standard datasets under severe occlusion, noise, and sparsity. Compared to the baseline DAPT, our method consistently produces more physically plausible skeleton structures that align better with the ground truth, especially in the highlighted occluded regions. These results demonstrate the superior robustness of our approach in handling highly degraded point cloud inputs.

As highlighted by the black ellipses, the baseline method (DAPT) exhibits significant performance degradation in these adverse conditions. For instance, in cases of extreme sparsity (e.g., at a LiDAR-to-subject distance of 26 m), DAPT often produces anatomically incorrect skeletons or distorted limb structures that fail to accurately represent the human pose. In contrast, our proposed method consistently produces more physically plausible skeletons that align much more closely with the ground truth (GT).

Specifically, even when large portions of the body are occluded or represented by only a few scattered points, our approach effectively captures the underlying geometric structure. These visualization results demonstrate the superior robustness of our method in handling highly degraded point cloud inputs compared to existing state-of-the-art baselines.

4.8. Computational efficiency

As shown in Table 7, QDRP, PointSP, APES, and SAMBLE have nearly identical parameter counts, FLOPs, and peak GPU memory, with differences of approximately 1% or less. QDRP introduces only a small latency overhead of about 1–2 ms per frame, while achieving the lowest MPJPE of 26.15 mm. These results demonstrate a favorable efficiency–performance trade-off for QDRP under a comparable computational budget.

5. Conclusion

In this work, we present QuanRA, a reliability-aware framework for LiDAR-based 3D human pose estimation that addresses the challenges of sparse and noisy point clouds. We introduce two key components: Multi-scale Region Structure Learning (MRSL), which effectively updates point features and assigns reliability scores to suppress noisy and irrelevant points, and Quantile-based Dynamic Region Partition (QDRP), which leverages these scores to dynamically organize reliable points into structured regions for robust pose modeling. Extensive experiments on LiDARHuman26M, SLOPER4D, Waymo, and HumanM3 demonstrate that our approach consistently achieves state-of-the-art performance across diverse scenarios. These results validate the effectiveness of reliability-aware feature aggregation and structured region formation in handling cluttered and irregular LiDAR point clouds, highlighting the potential of QuanRA for real-world 3D HPE applications.

CRediT authorship contribution statement

Tong Wu: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Zhongcheng Hong:** Writing – original draft, Visualization, Methodology. **Yanbin Liu:** Writing – review & editing, Supervision. **Sira Yongchareon:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] H. Akada, J. Wang, V. Golyanik, C. Theobalt, 3D human pose perception from egocentric stereo videos, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024, pp. 767–776.
[2] X. An, L. Zhao, C. Gong, J. Li, J. Yang, Pre-training a density-aware pose transformer for robust LiDAR-based 3D human pose estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2025, pp. 1755–1763.
[3] Y. Cai, W. Zhang, Y. Wu, C. Jin, Poseirm: enhance 3D human pose estimation on unseen camera settings via invariant risk minimization, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024, pp. 2124–2133.
[4] B. Chen, Y. Xia, Y. Zang, C. Wang, J. Li, Decoupled local aggregation for point cloud learning, arXiv preprint arXiv:2308.16532, 2023.
[5] S. Cheng, X. Chen, X. He, Z. Liu, X. Bai, Pra-net: point relation-aware network for 3D point cloud analysis, IEEE Trans. Image Process. 30 (2021) 4436–4448.
[6] A. Chharia, W. Gou, H. Dong, Mv-ssm: multi-view state space modeling for 3D human pose estimation, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2025, pp. 11590–11599.

[7] P. Cong, Y. Xu, Y. Ren, J. Zhang, L. Xu, J. Wang, J. Yu, Y. Ma, Weakly supervised 3D multi-person pose estimation for large-scale scenes based on monocular camera and single LiDAR, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 461–469.
[8] P. Cong, X. Zhu, F. Qiao, Y. Ren, X. Peng, Y. Hou, L. Xu, R. Yang, D. Manocha, Y. Ma, Stcrowd: a multimodal dataset for pedestrian perception in crowded scenes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19608–19617.
[9] J. Dai, H. Wang, Y. Zhao, Z. Liu, Apr-bica: LiDAR-based absolute pose regression with bidirectional cross attention and gating unit, Neurocomputing (2025) 132404.
[10] Y. Dai, Y. Lin, X. Lin, C. Wen, L. Xu, H. Yi, S. Shen, Y. Ma, C. Wang, Sloper4d: a scene-aware dataset for global 4D human pose estimation in urban environments, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 682–692.
[11] B. Fan, S. Wang, W. Guo, W. Zheng, J. Feng, J. Zhou, Human-m3: a multi-view multi-modal dataset for 3D human pose estimation in outdoor scenes, arXiv preprint arXiv:2308.00628, 2023.
[12] B. Fan, W. Zheng, J. Feng, J. Zhou, LiDAR-hmr: 3D human mesh recovery from LiDAR, IEEE Trans. Multimed. 27 (2025) 6962–6975.
[13] M. Fürst, S.T. Gupta, R. Schuster, O. Wasenmüller, D. Stricker, Hperl: 3D human pose estimation from RGB and LiDAR, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 7321–7327.
[14] M. Hasan, J. Hanawa, R. Goto, R. Suzuki, H. Fukuda, Y. Kuno, Y. Kobayashi, LiDAR-based detection, tracking, and property estimation: a contemporary review, Neurocomputing 506 (2022) 393–405.
[15] J. He, W. Yang, Video-based human pose regression via decoupled space-time aggregation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1022–1031.
[16] X. Hu, B. Zhong, Q. Liang, S. Zhang, N. Li, X. Li, Toward modalities correlation for RGB-T tracking, IEEE Trans. Circuits Syst. Video Technol. 34 (2024) 9102–9111.
[17] C. Li, P. Tang, X. Li, Y. Liu, T. Li, Enhancing sampling protocol for point cloud classification against corruptions, arXiv preprint arXiv:2408.12062, 2024.
[18] J. Li, J. Zhang, Z. Wang, S. Shen, C. Wen, Y. Ma, L. Xu, J. Yu, C. Wang, Lidarcap: long-range marker-less 3D human motion capture with LiDAR point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20502–20512.
[19] W. Li, M. Liu, H. Liu, P. Wang, J. Cai, N. Sebe, Hourglass tokenizer for efficient transformer-based 3D human pose estimation, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024, pp. 604–613.
[20] Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, S.T. Xia, Simcc: a simple coordinate classification perspective for human pose estimation, in: European Conference on Computer Vision, Springer, 2022, pp. 89–106.
[21] Z. Li, P. Gao, H. Yuan, R. Wei, Dynamic local feature aggregation for learning on point clouds, arXiv preprint arXiv:2301.02836, 2023.
[22] Z. Liao, J. Zhu, C. Wang, H. Hu, S.L. Waslander, Multiple view geometry transformers for 3D human pose estimation, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024, pp. 708–717.
[23] Z. Liu, H. Hu, Y. Cao, Z. Zhang, X. Tong, A closer look at local aggregation operators in point cloud analysis, in: European Conference on Computer Vision, Springer, 2020, pp. 326–342.
[24] G. Mei, L. Riz, Y. Wang, F. Poiesi, Geometrically-driven aggregation for zero-shot 3D point cloud understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27896–27905.
[25] J. Mei, A.Z. Zhu, X. Yan, H. Yan, S. Qiao, L.C. Chen, H. Kretzschmar, Waymo open dataset: panoramic video panoptic segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 53–72.
[26] M. Mohibullah, Y. Suda, Y. Hironaka, T. Miyawaki, R. Suzuki, M. Hasan, Y. Kobayashi, Privacy-preserving 3D human skeleton reconstruction from ankle-level 2D LiDAR using deep learning, Neurocomputing 131862 (2025).
[27] J. Peng, Y. Zhou, P. Mok, Ktpformer: kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024, pp. 1123–1132.
[28] Y. Ren, X. Han, Y. Yao, X. Long, Y. Sun, Y. Ma, Livehps + +: robust and coherent motion capture in dynamic free environment, in: European Conference on Computer Vision, Springer, 2024, pp. 127–144.
[29] Y. Ren, X. Han, C. Zhao, J. Wang, L. Xu, J. Yu, Y. Ma, Livehps: LiDAR-based scene-level human pose and shape estimation in free environment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1281–1291.
[30] A. Stathopoulos, L. Han, D. Metaxas, Score-guided diffusion for 3D human recovery, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024, pp. 906–915.
[31] C. Wang, M. Wu, S.K. Lam, X. Ning, S. Yu, R. Wang, W. Li, T. Srikanthan, Gpsformer: a global perception and local structure fitting-based transformer for point cloud understanding, in: European Conference on Computer Vision, Springer, 2024, pp. 75–92.
[32] J. Wang, J. Li, L. Ding, Y. Wang, T. Xu, Papooling: graph-based position adaptive aggregation of local geometry in point clouds, arXiv preprint arXiv:2111.14067, 2021.
[33] C. Wu, Y. Wan, H. Fu, J. Pfrommer, Z. Zhong, J. Zheng, J. Zhang, J. Beyerer, Samble: shape-specific point cloud sampling for an optimal trade-off between local detail and global uniformity, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2025, pp. 1342–1352.
[34] C. Wu, J. Zheng, J. Pfrommer, J. Beyerer, Attention-based point cloud edge sampling, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2023, pp. 5333–5343.
[35] X. Wu, L. Jiang, P.S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, H. Zhao, Point transformer v3: simpler faster stronger, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4840–4851.
[36] J. Xu, Y. Guo, Y. Peng, Finepose: fine-grained prompt-driven 3D human pose estimation via diffusion models, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024, pp. 561–570.
[37] M. Yan, X. Wang, Y. Dai, S. Shen, C. Wen, L. Xu, Y. Ma, C. Wang, Cimi4d: a large multimodal climbing motion dataset under human-scene interactions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12977–12988.
[38] X. Yan, C. Zheng, Z. Li, S. Wang, S. Cui, Pointasnl: robust point clouds processing using nonlocal neural networks with adaptive sampling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5589–5598.
[39] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C.X. Lu, L. Xie, Mm-fi: multi-modal non-intrusive 4D human dataset for versatile wireless sensing, Adv. Neural Inf. Process. Syst. 36 (2023) 18756–18768.
[40] D. Ye, Y. Xie, W. Chen, Z. Zhou, L. Ge, H. Foroosh, Lpformer: LiDAR pose estimation transformer with multi-task network, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 16432–16438.
[41] Z. Yu, Z. Sheng, Z. Zhou, L. Luo, S.Y. Cao, H. Gu, H. Zhang, H.L. Shen, Aggregating feature point cloud for depth completion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8732–8743.
[42] Z. Zeng, M. Dong, J. Zhou, H. Qiu, Z. Dong, M. Luo, B. Li, Deepla-net: very deep local aggregation networks for point cloud analysis, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 1330–1341.
[43] J. Zhang, Q. Mao, G. Hu, S. Shen, C. Wang, Neighborhood-enhanced 3D human pose estimation with monocular LiDAR in long-range outdoor scenes, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 7169–7177.
[44] Q. Zhao, C. Zheng, M. Liu, C. Chen, A single 2D pose with context is worth hundreds for 3D human pose estimation, Adv. Neural Inf. Process. Syst. 36 (2023) 27394–27413.
[45] J. Zheng, X. Shi, A. Gorban, J. Mao, Y. Song, C.R. Qi, T. Liu, V. Chari, A. Cornman, Y. Zhou, et al., Multi-modal 3D human pose estimation with 2D weak supervision in autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4478–4487.

Author biography

Tong Wu is currently a PhD candidate with the Department of Data Science and Artificial Intelligence, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology (AUT), Auckland, New Zealand.

Zhongcheng Hong is currently a PhD candidate with the Department of Data Science and Artificial Intelligence, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology (AUT), Auckland, New Zealand.

Yanbin Liu received the PhD degree from the Australian Artificial Intelligence Institute (AAII), University of Technology Sydney (UTS), Australia, under the supervision of Prof. Yi Yang. He is currently a Senior Lecturer with the Department of Data Science and Artificial Intelligence, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology (AUT), New Zealand. Before joining AUT, he was a Research Fellow at the School of Computing, Australian National University, and earlier held research fellow and postdoctoral positions at the University of Western Australia (jointly with the Harry Perkins Institute of Medical Research) and the University of Technology Sydney. His research interests include deep learning with limited labels and under distribution shifts, with applications to few-shot and meta-learning, spatio-temporal modeling, and medical imaging.

Sira Yongchareon received a PhD degree from Swinburne University of Technology, Melbourne, Australia. He is currently an Associate Professor with the Department of Data Science and Artificial Intelligence, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology (AUT), Auckland, New Zealand. His research interests include pervasive computing, AI/deep learning and data management for the Internet of Things (IoT), activity recognition, wireless sensing, and mobile/edge computing in intelligent environments such as smart cities, smart homes, transportation, and healthcare. He is a Senior Member of IEEE and a Senior Member of ACM.