Diffusion Model is a Good Pose Estimator from 3D RF-Vision

Junqiao Fan[1], Jianfei Yang[1,2*], Yuecong Xu[1], and Lihua Xie[1]

[1] School of Electrical and Electronic Engineering
[2] School of Mechanical and Aerospace Engineering
Nanyang Technological University, Singapore
{fanj0019, jianfei.yang, xuyu0014, elhxie}@ntu.edu.sg

Abstract. Human pose estimation (HPE) from Radio Frequency vision (RF-vision) performs human sensing using RF signals that penetrate obstacles without revealing privacy (e.g., facial information). Recently, mmWave radar has emerged as a promising RF-vision sensor, providing radar point clouds by processing RF signals. However, the mmWave radar has a limited resolution with severe noise, leading to inaccurate and inconsistent human pose estimation. This work proposes mmDiff, a novel diffusion-based pose estimator tailored for noisy radar data. Our approach aims to provide reliable guidance as conditions to diffusion models. Two key challenges are addressed by mmDiff: (1) miss-detection of parts of human bodies, which is addressed by a module that isolates feature extraction from different body parts, and (2) signal inconsistency due to environmental interference, which is tackled by incorporating prior knowledge of body structure and motion. Several modules are designed to achieve these goals, whose features work as the conditions for the subsequent diffusion model, eliminating the miss-detection and instability of HPE based on RF-vision. Extensive experiments demonstrate that mmDiff outperforms existing methods significantly, achieving state-of-the-art performances on public datasets. [1]

1 Introduction

Human pose estimation (HPE) has been a widely-studied computer vision task for predicting coordinates of human keypoints and generating human skeletons [3, 24, 53]. It is a fundamental task for human-centered applications, such as augmented/virtual reality [25, 45, 52], rehabilitation [4, 38], and human-robot interaction [10, 13]. Current HPE solutions mainly rely on RGB(D) cameras [9, 20]. Though demonstrating promising accuracy, camera-based pose estimators have intrinsic limitations under adverse environments, e.g., smoke, low illumination, and occlusion [41]. The privacy issue of cameras also hinders the viability of HPE in medical scenarios, e.g., rehabilitation systems in hospitals [1].

* J. Yang is the project lead.
[1] The project page of mmDiff is https://fanjunqiao.github.io/mmDiff-site/.

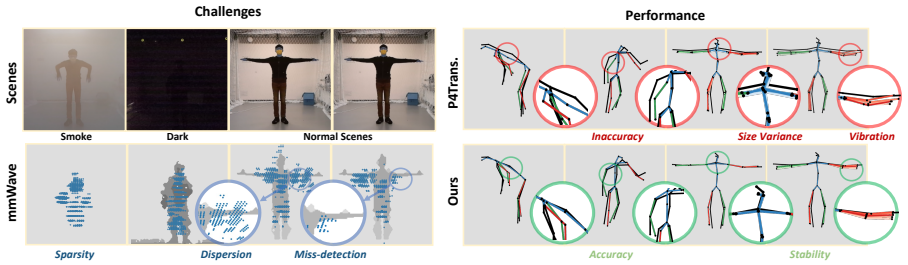

Fig. 1: Left: challenges of mmWave PCs. Right: the performance of existing SOTA (P4Transformer [12]) compared to ours. The GTs are black and predictions are colored. PC's sparsity and dispersion cause inaccurate spline and shoulder. Inconsistent PCs with occasional miss-detection further cause size variance and pose vibration. mmDiff proposes diffusion-based pose estimation with enhanced accuracy and stability.

Overcoming limitations of camera-based HPE, Radio-Frequency vision (RF-vision) has attracted surging attention for human sensing. The emerging mmWave radar technology presents a promising and feasible solution for HPE due to its price, portability, and energy efficiency [40]. Operated at the frequencies of 30-300 GHz [46], commercial mmWave radars transmit and receive RF signals that penetrate human targets and occlusions. Radar point clouds (PCs) are further extracted as the salient target detection via monitoring signals' characteristic changes [30, 43]. Therefore, the extracted PCs are more robust to adverse environments [48] and reveal little privacy [44], inspiring accurate and privacy-preserving HPE [2, 22, 43, 44]. However, due to the bandwidth and hardware limitations [26], radar PCs are sparse with limited geometric information [2], leading to huge difficulties in achieving HPE. The sparse PCs are noisy in two aspects: (1) mmWave radar has a lower spatial resolution, leading to PC dispersion throughout the target area accompanied by ghost points caused by multi-path effect [26, 37]; (2) signal's specular reflection and interference [5, 6] further cause inconsistent sensing data, leading to occasional miss-detection of human parts.

To deal with sparse and noisy mmWave PCs, existing solutions mainly rely on kinds of data augmentations, e.g., multi-frame aggregation to enhance PC resolution [2,43]. For the feature extractor, they directly borrow Long Short-Term Memory (LSTM) [43] or transformer-based architectures [6, 44] from existing RGB(D) HPE methods. However, these feature encoders are tailored for visual and language modalities, which struggle to handle noisy and inconsistent radar PCs [27]. As shown in Figure 1 (right), the existing SOTA solution [12] still suffers from pose vibration and severe drift, achieving undesirable performance.

Denoising Diffusion Probabilistic Models (DDPMs) [15, 34], also known as diffusion models, have demonstrated superior performance in various generative tasks, such as image generation and image restoration [14, 16, 54]. Diffusion models perform progressive noise elimination, transferring noisy distribution into desired target distribution [35]. Inspired by such capability, we aim to mitigate the noise of mmWave HPE, which motivates mmDiff, a diffusion-based pose

estimator tailored for noisy radar PCs. Different from existing diffusion-based HPE using RGB(D), HPE using mmWave PCs confronts two key challenges: (1) extracting robust features from noisy PCs where miss-detection of human bodies may happen, and (2) overcoming signal inconsistency for stable HPE. For the first challenge, we propose to isolate the feature extraction for different body joints, so that occasional miss-detection would not affect the feature extraction of detected joints. Extracting features directly from local PCs also improves the feature resolution. For the inconsistency issue, prior knowledge of human body structure and motion can reduce unreasonable cases, achieving consistent feature learning. As the human structure has a size constraint where limb-length should remain constant [8], the limb-length can be additionally estimated to prevent pose variance. Moreover, inspired by human motion continuity [29] that discourages abrupt human behavior changes, historical poses can be leveraged for pose generation refinement, minimizing pose vibration.

To this end, mmDiff first designs a conditional diffusion model capable of injecting radar information as the guiding conditions. Four modules are designed to extract clean and consistent information from radar point clouds: (1) Global radar context is proposed to isolate the globally extracted features for different human joints using a transformer [39], which generates more robust joint-wise features to handle miss-detection. (2) Local radar context is proposed to extract local features around body joints with a local transformer, which performs point-level attention for higher resolution. (3) Structural human limb-length consistency is proposed to extract human limb-length as consistent patterns, which reduce limb-length variance. (4) Temporal motion consistency is proposed to learn smooth motion patterns from historical estimated poses, which avoid pose vibration. Experiments have shown a significant improvement in pose estimation accuracy using mmDiff compared to the state-of-the-art models. Meanwhile, the generated poses demonstrate comparable structural and motion stability, validating the effectiveness of our designed conditional modules.

In summary, our contributions are three-fold. First, we propose a novel diffusion-based HPE framework with sparse and noisy mmWave radar PCs. To the best of our knowledge, mmDiff is the first diffusion-based paradigm for mmWave radar-based HPE. Second, four modules are proposed to extract robust representations from the noisy and inconsistent radar PCs considering global radar context, local radar context, structural limb-length consistency, and temporal motion consistency, used as the conditions to guide the diffusion process. Finally, extensive experiments show our approach achieves state-of-the-art performance on two public datasets: mmBody [6] and mm-Fi [44].

2 Related Work

mmWave Human Pose Estimation. For decades, extensive works [3, 24] have been focused on human pose estimation from RGB(D) images. Though achieving desirable accuracy, the major challenge faced by RGB(D) HPE is the performance drop under adverse environments and with self-occlusion [7]. Recently, commercial

mmWave radar has been proven to extract sufficient information for human body reconstruction [43], bringing the potential for mmWave HPE. Despite methods [21,42] utilizing raw radar signals for mmWave HPE, point clouds-based methods [1,2,6,43,44] become popular for its format uniformity. Particularly, Xue et al. [43] utilizes the LSTM model with anchor-based local encoding to deal with the noisy nature of radar. An et al. [2] integrate point clouds of consecutive frames to handle the point cloud sparsity. Chen et al. [6] and Yang et al. [44] propose transformer-based benchmarks based on their dataset. However, the sparse and noisy radar point clouds still hinder the accuracy of HPE using existing non-parametric regression models.

Diffusion Model for Human Pose Estimation. Diffusion models have been widely applied in image generation, such as image restoration [23], super-resolution [18], and text-to-image synthesis [31]. Since the proposition of the denoising diffusion probabilistic model (DDPM) [15], the diffusion model is extended to a wider range of generative applications, including pose/skeleton generation. DiffPose [14] formulate the 3D pose estimation problem as a pose generation task from low-determinant 2D poses to high-determinant 3D poses. Following DiffPose, Shan et al. [33] proposes to generate multi-hypothesis poses for 3D pose ambiguity and Saadatnejad et al. [32] proposes to predict the human poses in future time frames. Meanwhile, RGB-guided diffusion models for 2D and 3D human pose generation are also achieved [28,51]. Nevertheless, existing methods focus on diffusion-based HPE from stable and informative modality sources, either well-estimated 2D poses or high-resolution RGB images. None of these works investigate how noisy and sparse mmWave radar point clouds are used to guide the pose generation.

3 Methodology

We address 3D HPE task with noisy and sparse PCs from mmWave radar. At each time frame $t \geq 0$, given the input 6-dimensional radar point cloud $R_t \in \mathbb{R}^{N\times 6}$, where N denotes the number of detected points, our task is to estimate the ground-truth 17-joint human pose $H_t = \{h_t^1, h_t^2, ..., h_t^{17}\} \in \mathbb{R}^{17\times 3}$. Apart from the Cartesian coordinate $\{x, y, z\}$, each detected radar point contains three attributes $\{v, E, A\}$, representing velocity, energy, and amplitude respectively.

To extract clean radar features and overcome mmWave signal inconsistency, we propose mmDiff as a diffusion-based paradigm for mmWave radar-based HPE. The overview of the architecture of mmDiff is presented in Figure 2. We start with an overall description of how human poses are generated by diffusion models using radar information as the guiding conditions in Section 3.1, followed by a detailed illustration of the four modules for extracting robust representations from the noisy and inconsistent radar PCs in Section 3.2 and Section 3.3. We omit the subscript t denoting the time frame in subsequent discussions for clarity.

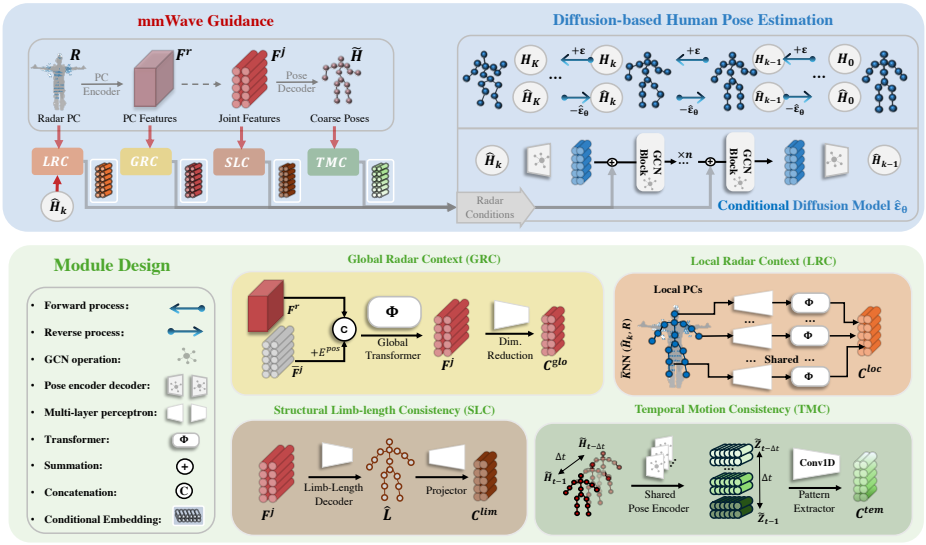

Fig. 2: mmDiff proposes a diffusion-based HPE model, using mmWave radar information as conditions. $k \in [0..K]$ denotes the diffusion step. Four modules are proposed as the more reliable guidance, addressing PCs' noise and inconsistency: GRC and LRC first extract robust global-local radar features, C^{glo} and C^{loc}; SLC and TMC then extract consistent human structure and motion patterns, C^{tem} and C^{lim}.

3.1 Diffusion-based Human Pose Estimation

Diffusion models [15, 35] are a category of probabilistic generative models popular in various tasks, e.g., image generation. Given a noisy image from a noisy distribution, diffusion models can generate realistic image samples that match the natural image distribution, through iterative noise removal [18, 23]. Extended to HPE, diffusion models can estimate the distribution of reasonable human poses for realistic pose generation. Particularly with noisy radar modality, miss-detected body joints can be accurately estimated by inferring from the detected ones, and inaccurate joints causing twisted human structures are potentially refined.

The diffusion-based HPE consists of two processes: (1) a forward process gradually generates noisier samples as the training guidance and (2) a reverse process learns to invert the forward diffusion. Modelled by a Markov chain of length K, the forward process starts from the ground truth pose H_0, iteratively samples a noisier pose H_k by adding Gaussian noise $\epsilon \in \mathcal{N}(0, I)$:

$$q\left(H_k \mid H_{k-1}\right) = \mathcal{N}\left(H_k \mid \sqrt{1-\beta_k} H_{k-1}, \beta_k I\right), \tag{1}$$

where $k \in [0..K]$ refers to the diffusion step and β_k refers to the noise scale. On the contrary, the reverse process starts from a noisy pose initialization $\hat{H}_K$ and progressively removes noises until $\hat{H}_0$ is generated. For each step k, a diffusion model $\hat{\varepsilon}_\theta$ is trained to identify the pose noise $\hat{\varepsilon}_k$ within $\hat{H}_k$, and remove it for

more reliable pose $\hat{H}_{k-1}$:

$$\hat{\varepsilon}_k = \hat{\varepsilon}_\theta(\hat{H}_k, k), \tag{2}$$

$$\hat{H}_{k-1} = (1-\beta_k)^{-1}(\hat{H}_k - \beta_k \hat{\varepsilon}_k). \tag{3}$$

To better leverage the context information provided by radar's sensing data, we propose a set of conditions C containing latent feature embeddings extracted from the mmWave modality, to guide each step of the reverse process:

$$\hat{\varepsilon}_k = \hat{\varepsilon}_\theta(\hat{H}_k, C, k). \tag{4}$$

As an implementation, we propose a conditional diffusion model, which injects the radar conditions C in the latent feature space, using a Graph Convolution Network (GCN) [49] as the backbone. The GCN takes the 17-joint human pose $H_k \in \mathbb{R}^{17\times 3}$ as input, which is subsequently encoded into the latent pose embedding $Z_k \in \mathbb{R}^{17\times 96}$ by a GCN encoder, fed into n GCN blocks, and decoded into $\hat{\varepsilon}_k \in \mathbb{R}^{17\times 3}$ by a GCN decoder. To inject radar conditions, we propose to add the conditional embeddings from C to the pose feature Z_k before each GCN block, which serves as extra information for more accurate noise estimation of $\hat{\varepsilon}_k$. To align features, the conditional embeddings are also projected to the $\mathbb{R}^{17\times 96}$ feature space. In the following sections, we discuss in detail how to extract clean and consistent radar features that construct C.

3.2 Global-local Radar Context

Accurate guidance for the diffusion model depends on robust radar feature extraction, which should carefully handle the miss-detection of human bodies. In this section, we first revisit the existing mmWave feature extraction paradigm. Then, we further propose to improve it with two modules: (1) Global Radar Context (GRC) extracting features from overall PCs that can handle miss-detection; and (2) Local Radar Context (LRC) extracting local PC features near body joints for higher resolution.

Revisit mmWave Feature Extraction. Existing mmWave HPE paradigm applies encoder-decoder architectures to encode the radar PCs, $R \in \mathbb{R}^{N\times 6}$, into latent feature representations and decode them into human poses. Anchor-based methods [12, 47] are common options for PC encoders, where point anchors are first sampled from the Farthest Point Sampling algorithm and nearby PCs are extracted as the anchor features. As a result, the holistic radar PCs are encoded into PC features $F^r \in \mathbb{R}^{P\times 1024}$, where P indicates the anchor number. To decode the PC features, a Multi-Layer Perceptron (MLP) is commonly applied as a dimension-reducing projection to generate the joint feature $F^j \in \mathbb{R}^{17\times 1024}$, which is further decoded into human poses $\tilde{H} \in \mathbb{R}^{17\times 3}$ by another MLP. However, occasional radar miss-detection causes uncertainty within the PC features F^r, while MLP-based projection can hardly identify miss-detected joints. Additionally, radar low-resolution PCs further impose noises. As a result, the estimated human joints are generally coarse.

Global Radar Context (GRC). To handle occasional miss-detection, GRC is proposed to isolate feature extraction for different body joints, using global information from F^r to construct more robust joint features F^j. From joint features, the diffusion model potentially identifies miss-detected joints and utilizes human prior knowledge for more accurate estimation. We exploit a Global-Transformer Φ^g to facilitate joint-wise feature extraction from the F^r. Following the *cls* token design in ViT [11], we first randomly initialize a trainable joint feature template (with no information) $\bar{F}^j \in \mathbb{R}^{17\times 1024}$ and add it with positional embedding $E^{pos} \in \mathbb{R}^{17\times 1024}$. Then, the joint feature template is concatenated with the PC feature $[\bar{F}^j, F^r] \in \mathbb{R}^{(17+P)\times 1024}$ and fed into Φ^g:

$$F^j, F'^r = \Phi^g(\bar{F}^j, F^r). \tag{5}$$

The output $F^j \in \mathbb{R}^{17\times 1024}$ is selected as the joint feature and the rest F'^r is ignored. Within the transformer, deep correlation is captured, not only within F^r but also between F^j and F^r. Each body joint performs individual feature extraction based on their correlation with the PC feature, so that features of detected joints are less affected by other undetected parts. Finally, as the 1024-dim F^j feature space is too sparse for the diffusion latent condition, an MLP-based dimension-reduction function g^g further condenses the extracted information within a $\mathbb{R}^{17\times 64}$ conditional embedding:

$$C^{glo} = g^g(F^j). \tag{6}$$

Local Radar Context (LRC). LRC further performs local point-to-point self-attention to extract higher-resolution features from local PCs near body joints. To select local PCs, existing methods [43] utilize static joint anchors from coarsely-estimated human poses $\tilde{H}$. However, the local joint features should dynamically reflect the joints' errors at different diffusion steps. Therefore, we propose dynamic joint anchors from intermediate diffusion poses $\hat{H}_k$ for PC selection. With $i \in [1, .., 17]$ and each joint $\hat{h}^i_k$ from the $\hat{H}_k$ as an anchor, the $\bar{K}$-nearest-neighbors ($\bar{K}$NN) algorithm is applied to select $\bar{K}$ nearest points as local PCs, $\bar{R}^i_k \in \mathbb{R}^{\bar{K}\times 6}$. Each $\bar{R}^i_k$ is first encoded by a shared MLP g^l into a $\mathbb{R}^{\bar{K}\times 64}$ embedding, and then fed into a shared small-scale local transformer Φ^l for point-to-point self-attention. Finally, average pooling is performed to generate $\mathbb{R}^{64}$ embeddings, which are further aggregated (concatenated) for every joint anchor $\tilde{h}^i$ as the conditional embedding:

$$C^{loc} = \bigcup_{i\in[1,..,17]} \Phi^l \circ g^l(\bar{R}^i_k). \tag{7}$$

3.3 Structural-motion Consistent Patterns

Inconsistent radar signals such as occasional miss-detection lead to discontinuous and unstable pose estimation, such as variant limb-length, pose vibration, or inconsistent error frames. To mitigate such inconsistency, we further extract

consistent human patterns based on human structure and motion prior knowledge: (1) Structural Limb-length Consistency (SLC) that extracts limb-length patterns to reduce limb-length variance, and (2) Temporal Motion Consistency (TMC) learns smooth motion patterns from historically estimated human poses.

Structural Limb-Length Consistency (SLC). SLC learns to extract a human-size indicator, the 16 limb-length $\hat{L} = \{\hat{l}^1, ..., \hat{l}^{16}\} \in \mathbb{R}^{16}$, where each limb-length measures the bone length connecting adjacent body joints. The extracted limb-length serves as a structural constraint to reduce the limb-length variance during the pose generation. Similar to decoding a pose, an MLP-based limb decoder g_1^{lim} is first applied to decode the previously extracted global joint feature F^j into predicted limb-length $\hat{L}$. Though with less dimension, the extracted limb-length contains physical meanings and thus is more consistent and stable. To ensure accurate limb-length decoding, a limb loss is designed to guide the training (details in Section 3.4). To further project the estimated limb-length $\hat{L}$ into latent feature space, another MLP-based projector g_2^{lim} then transforms the $\hat{L}$ into the $\mathbb{R}^{96}$ embedding, which is further broadcasted to different joints as the $\mathbb{R}^{17\times96}$ conditional embedding:

$$C^{lim} = g_2^{lim}(\hat{L}) = g_2^{lim} \circ g_1^{lim}(F^j). \tag{8}$$

Temporal Motion Consistency (TMC). Inspired by the fact that human motion is generally stable and consistent [29], multi-frame historical human poses can be utilized to extract the motion patterns in estimating the current pose. Such motion patterns provide temporal constraints for the diffusion model, which avoids error frames and pose vibration. As shown in Figure 2, TMC extracts the latent motion patterns from a sequence of historical-estimated coarse poses $\{\tilde{H}_{t-\Delta t}, ..., \tilde{H}_{t-1}\}$, where Δt is the number of historical frames. Firstly, a shared GCN encoder g_1^{tem} is applied to convert the pose sequence into feature embedding sequence $\{\tilde{Z}_{(t-\Delta t)}, ..., \tilde{Z}_{(t-1)}\} \in \mathbb{R}^{\Delta t\times(17\times96)}$. Then, a 1D-convolution-based pattern extractor g_2^{tem} is applied to extract the motion information along the temporal dimension, which generates a $\mathbb{R}^{17\times96}$ temporal embedding as C^{tem}:

$$C^{tem} = g_2^{tem}\left(\bigcup_{i\in[1..\Delta t]} \tilde{Z}_{(t-i)}\right) = g_2^{tem}\left(\bigcup_{i\in[1..\Delta t]} g_1^{tem}(\tilde{H}_{t-i})\right). \tag{9}$$

The reason for using 1D convolution is two-fold: (1) smooth motion features can be extracted by potentially averaging pose features of historical frames, which avoids pose vibration; and (2) the motion trend of the on-performing actions is potentially extracted to avoid inconsistent error frames. For example, increasing z values of the hand's location are expected when performing the 'raising hand'.

3.4 Overall Learning Objective

As feature extraction from global radar PCs is computation-consuming, we divide the training process into two phases. Phase one facilitates the extraction of

the global joint features F^j and coarse estimation of human poses $\tilde{H}$. The GRC, PC encoder (from an off-the-shelf mmWave HPE network), and an MLP-based pose decoder are trained together. The learning objective of the phase one is minimizing the $\mathcal{L}_2$ pose regression loss $\mathcal{L}_{\text{joint}}$:

$$\mathcal{L}_{\text{joint}} = E_{i\sim[1,17]}||h^i - \tilde{h}^i||_2^2. \tag{10}$$

In phase two, the remaining three conditional modules and the diffusion model are trained together, with the phase one parameters frozen. The extracted F^j and $\tilde{H}$ serve as the input of structural-motion consistency modules, and $\tilde{H}$ initialize $\hat{H}_K$ for the reverse diffusion process. The learning objective is minimizing $\mathcal{L}_{\text{diff}}$, which is the diffusion learning objective following DDPM [15]. To ensure accurate limb-length estimation, an $\mathcal{L}_1$ limb regression loss is further designed and integrated:

$$\begin{aligned}\mathcal{L}_{\text{diff}} = \mathbb{E}_{k\sim[1,T]}\mathbb{E}_{\varepsilon_k\sim\mathcal{N}(0,I)}||\varepsilon_k - \hat{\varepsilon}_\theta(H_k, k, C)||_2^2 \\ +\lambda * \mathbb{E}_{i\sim[1,16]}|l^i - \hat{l}^i|_1,\end{aligned} \tag{11}$$

where λ is a weighting parameter and each l^i is the ground truth limb-length calculated from ground truth pose H.

4 Experiments

4.1 Experiment Setup

Datasets. mmBody [44] studies the robustness of human sensing with various sensors: RGB, Depth, and mmWave radar. Human skeletons are annotated by the MoCap system. Models are set to train on data collected from 2 standard scenes (Lab1 and Lab2), and tested on 3 basic scenes and 4 adverse scenes (including unseen subjects). Meanwhile, mm-Fi [44] is a larger-scale HPE dataset using a lower-bandwidth mmWave radar with sparser PCs. The skeleton annotations are rather unstable compared to the MoCap annotations, as obtained by RGB using the pretrained HRNet-w48 [36]. Three data-splitting methods are used with a train-test split ratio of 4:1: random, cross-subject, and cross-environment.

Implementation Details. Our methods are trained for 100 epochs with a batch size of 1024. The Adam algorithm [19] is used for optimization, with the learning rate as $2e-5$, the gradient clip as 1.0, and the momentum as 0.9. The forward/reverse diffusion process is set as $K = 25$ steps with a constant β sampling of 0.001. An average of 5 hypotheses is recorded for a fair comparison with non-diffusion methods. We choose Point4D [12] as the GRC's PC encoder for mmBody and PointTransformer [47] for mm-Fi. We set $\bar{K} = 50$ for $\bar{K}$NN algorithm for LRC, but due to insufficient radar points ($N < 100$), the LRC module is neglected for mm-Fi. We further set $\Delta t = 8$ for TMC and $\lambda = 5$ for SLC. The hyper-parameters are obtained empirically, more precise hyper-parameter tuning tricks such as the Bayesian optimization could lead to better results.

Table 1: Quantitative results on mmBody [6], evaluated by MPJPE in white and PA-MPJPE in grey. * indicates diffusion-based methods. G, L, T, and S denote GRC, LRC, TMC, and SLC respectively. Bold is the best.

Methods	Basic Scenes						Adverse Environment								Average	
	Lab1		Lab2		Furnished		Rain		Smoke		Dark		Occlusion			
RGB [6]	74	/	73	/	71	/	80	/	86	/	105	/	/	/	81	/
Depth [6]	55	/	39	/	55	/	86	/	243	/	51	/	/	/	88	/
RGB(D) [6]	58	/	34	/	54	/	95	/	154	/	58	/	/	/	75	/
MM-Mesh [43]	95.1	69.48	87.87	77.3	93.28	73.14	106.9	72.38	106.7	76.52	83.54	64.19	85.55	62.5	94.13	70.79
P4transformer [12]	69.35	54.45	73.40	66.39	75.28	55.77	86.83	65.71	89.82	69.73	73.48	54.52	78.56	57.36	78.10	60.56
PoseFormer [50]	64.53	51.61	70.17	63.26	69.71	51.04	77.49	59.03	84.82	63.57	69.88	50.46	73.52	53.53	72.87	56.07
DiffPose* [14]	66.43	52.56	68.36	65.69	69.78	51.29	77.77	62.60	89.01	69.34	67.27	49.90	74.52	56.20	73.31	58.23
mmDiff(G)*	61.00	49.19	**67.79**	62.45	69.83	51.47	77.39	60.48	81.41	64.44	68.83	49.82	70.64	52.79	70.99	55.80
mmDiff(G,L)*	61.11	49.41	69.06	63.14	68.17	50.60	73.70	58.45	82.26	64.30	66.06	48.56	68.18	50.8	69.79	55.04
mmDiff(G,L,T)*	59.90	**47.81**	68.12	62.02	**66.98**	**48.63**	74.84	**58.13**	80.95	63.44	65.25	**47.26**	68.05	50.4	69.16	53.96
mmDiff(G,L,T,S)*	**59.52**	47.85	69.36	**61.23**	67.15	49.19	**71.03**	58.40	**76.92**	**62.25**	**65.08**	47.47	**67.47**	**49.54**	**68.08**	**53.71**

Table 2: Quantitative results on mm-Fi [44].

Methods	Random		Cross-Subject		Cross-Environment	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
PointTransformer [47]	$73.09_{\pm 2.70}$	$55.60_{\pm 1.40}$	$75.96_{\pm 6.90}$	$58.70_{\pm 4.30}$	$88.28_{\pm 4.50}$	$68.79_{\pm 2.79}$
Diffpose* [14]	$73.44_{\pm 0.29}$	$56.83_{\pm 0.25}$	$70.31_{\pm 0.27}$	$54.12_{\pm 0.31}$	$86.35_{\pm 0.06}$	$66.87_{\pm 0.17}$
mmDiff(G)*	$68.62_{\pm 0.06}$	$53.11_{\pm 0.05}$	$68.46_{\pm 0.06}$	$52.55_{\pm 0.05}$	$85.63_{\pm 0.53}$	$66.43_{\pm 0.28}$
mmDiff(G,S)*	$65.72_{\pm 0.08}$	$50.72_{\pm 0.01}$	$67.18_{\pm 0.18}$	$51.85_{\pm 0.05}$	$83.39_{\pm 0.17}$	$64.61_{\pm 0.40}$
mmDiff(G,S,T)*	$\mathbf{65.26_{\pm 0.11}}$	$\mathbf{50.35_{\pm 0.09}}$	$\mathbf{65.62_{\pm 0.24}}$	$\mathbf{50.23_{\pm 0.24}}$	$\mathbf{82.73_{\pm 0.62}}$	$\mathbf{63.87_{\pm 0.26}}$

Compared Methods. (1) RGB, Depth, and RGB(D) [6] are benchmarks using different modalities. (2) mmWave methods: P4Transformer [12] is the benchmark on mmBody containing Point4D Convolution as the PC encoder and a transformer for self-attention; PointTransformer [47] is the benchmark on mm-Fi designed to handle the sparser PCs. mmMesh [43] is implemented as an extra mmWave HPE baseline on mmBody. (3) SOTA camera-based HPE methods: camera-based 3D HPE methods generally perform pose lifting from 2D poses to 3D poses. To modify them to perform radar-based HPE, we train the models to perform pose refinements, from coarse poses $\tilde{H}$ to clean 3D poses. PoseFormer [50] is the transformer-based method with temporal-spatial attention. DiffPose [14] is the diffusion-based method using SOTA graph-based GraFormer [49] as backbones.

Evaluation Metric. Two evaluation metrics are adopted following [17]: (1) Mean Per Joint Position Error (MPJPE (mm)): the average joint error between ground truth and prediction (after pelvis alignment); and (2) Procrustes Analysis MPJPE (PA-MPJPE (mm)): procrustes methods (translation, rotation, and scaling) are performed before error calculation.

4.2 Overall Result

Performance on mmBody. As shown in Table 1, mmWave-based methods have better robustness for cross-domain scenes and adverse environments. Particularly, our proposed mmDiff demonstrates superior results compared to related methods

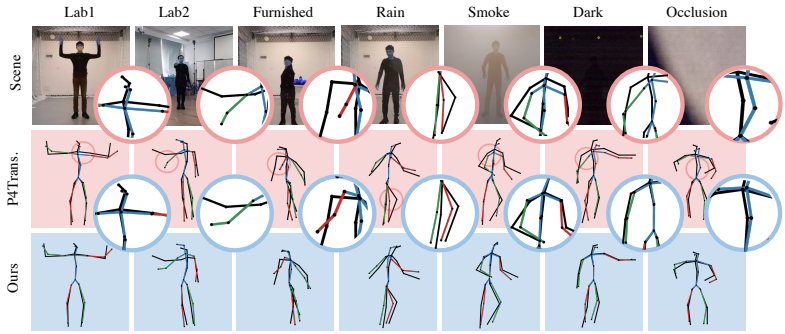

Fig. 3: Qualitative visualization of the estimated poses on mmBody dataset. mmDiff demonstrates higher keypoint accuracy. The GTs are black and predictions are colored.

on the mmBody dataset. Compared to the SOTA mmWave HPE method, our method mmDiff(G,L,T,S) outperforms the P4Transformer [12] by 12.8% (MPJPE) and 11.3% (PA-MPJPE). For adverse environments, the improvement is more significant with 14.7% (MPJPE) and 12.0% (PA-MPJPE), because radar signals get noisier due to the specular reflection and mmDiff has improved noise-handling capability. Compared to Diffpose [14] and PoseFormer [50] that utilize SOTA 3D HPE methods for pose refinement, mmDiff(G,L,T,S) still outperforms by 6.6% (MPJPE) and 4.2% (PA-MPJPE), demonstrating the proposed modules are dedicated to handle noisy and sparse radar modalities. Furthermore, mmDiff enables better mmWave-based performance compared to RGB-based methods under all scenes and RGB(D)-based methods under adverse environments. Qualitatively, we observe more accurate human poses as illustrated in Figure 3.

Performance on mm-Fi. As shown in Table 2, the generalizability of our proposed mmDiff is further explored on mm-Fi dataset. We compare the proposed method with the benchmark PointTransformer [47] and the SOTA DiffPose pose refinement [14]. Compared with the benchmark, our proposed method reduces the pose estimation error by 6.29% to 13.61% (MPJPE) and 7.15% to 14.42% (PA-MPJPE). Though the vanilla DiffPose for pose refinement can improve

Table 3: Ablation studies of proposed modules on mmBody.

		Diffusion Model		Context Modules			Consistency Modules			Modules Elimination				Overall
Modules	Diffusion		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Global			✓		✓					✓	✓	✓	✓
	Local				✓	✓				✓		✓	✓	✓
	Limb						✓		✓	✓	✓		✓	✓
	Temporal							✓	✓	✓	✓	✓		✓
Average	MPJPE	78.10	73.31	70.99	70.43	69.79	69.85	71.71	69.46	68.83	68.81	69.45	69.16	**68.08**
	PA-MPJPE	60.56	58.23	55.80	55.45	55.04	54.90	57.06	54.70	54.18	54.58	54.53	53.96	**53.71**

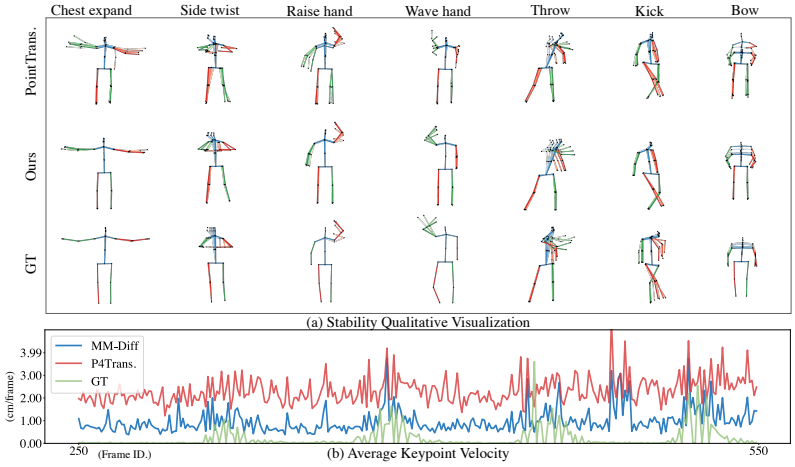

Fig. 4: (a) shows the pose motion stability on mm-Fi by plotting 5 consecutive frames of poses. mmDiff shows more consistent motion patterns (zoom in for details). (b) shows the motion energy levels on mmBody, where lower AKV indicates better stability.

the result, the performance is further improved by 4.19% to 11.14% (MPJPE) and 4.49% to 11.40% (PA-MPJPE) with the integration of the radar context module and the consistency modules, further demonstrating their effectiveness. Moreover, using cross-subject splitting, our method shows comparable results to random splitting. As our model explicitly learns the human structural and motion patterns from the radar signals, these patterns potentially generalize for unseen subjects. Our method also demonstrates cross-domain ability as the performance steadily improves with the integration of different modules.

4.3 Analytics

Ablation Study. As shown in Table 3, we perform the ablation study to verify the effectiveness of each component from the following 4 perspectives. (1) The vanilla diffusion design without any radar guidance can reduce the joint error by modeling pose distribution and refine deformed poses. (2) The effectiveness of the global-local radar context modules is verified by the performance gain of 4.8% (MPJPE) and 5.5% (PA-MPJPE). Both modules extract clean radar

Table 4: Ablation studies of the detailed designs in global-local radar context modules.

Global Radar Context				
Modules	Diffusion	✓	✓	✓
	PC Features F^r		✓	
	Joint Features F^j			✓
Average	MPJPE	73.31	72.50	**70.99**
	PA-MPJPE	58.23	58.03	**55.80**

Local Radar Context				
Modules	Diffusion	✓	✓	✓
	Static Anchors		✓	
	Dynamic Anchors			✓
Average	MPJPE	73.31	71.05	**70.43**
	PA-MPJPE	58.23	56.06	**55.45**

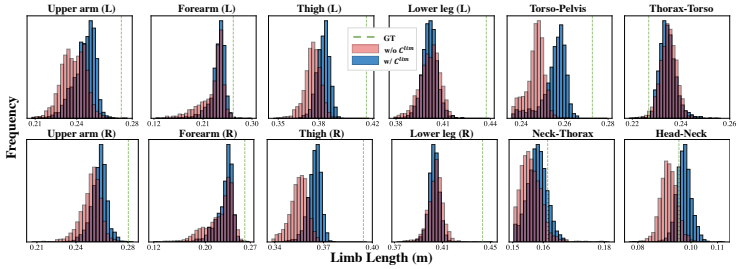

Fig. 5: Limb-length distribution for a single subject by histogram. The error within 5 cm can be treated as correct. With C^{lim}, more accurate limb-length and less variance are observed, as the distribution moves towards the GT and is more concentrated.

information with a different focus (local radar PCs or global PC features). (3) The effectiveness of the structure-motion consistency modules are verified by the performance gain of 5.3% (MPJPE) and 6.1% (PA-MPJPE). (4) The necessity of each module is further proved, as the elimination of different modules leads to a performance drop. Specifically, the performance drop is more significant when removing the limb-length module (C^{lim}) and temporal motion consistency module (C^{tem}), as the pose instability causes great pose inaccuracy.

Effect of Global-local Radar Context. To demonstrate the effectiveness of joint-wise feature extraction using the joint feature template, we compare our design with an alternative in Table 4: to directly generate PC feature guidance using the transformer without any template. Our designed joint feature guidance significantly outperforms the PC feature guidance, as the PC features are easily affected by the radar's miss-detection. To demonstrate the effectiveness of dynamic anchors of LRC, in Table 4 we compare our design with static anchor design [43] using coarse estimated human poses $\tilde{H}$. Our design has better performance and is more suitable for diffusion-based local PC selection.

Effects of Temporal Motion Consistency. In Figure 4 (a), we observe a mixing of incorrect skeletons within the correct skeleton timeline with PointTrans. [47], e.g. chest-expanding and hand-waving. With mmDiff, smooth motion patterns are ensured based on human behavioral prior knowledge. Additionally, our proposed mmDiff can mitigate the uncertain locations of legs and arms (caused by low resolution), demonstrating enhanced pose accuracy. Furthermore, for throwing and kicking actions, our method is more stable compared to RGB-extracted ground truth, as camera-based HPE suffers from self-occlusion. In Figure 4 (b), we further apply the Average Keypoint Velocity (AKV) to quantitatively measure the pose stability. AKV is defined as $E_{i=[0..17]}(||J_t^i - J_{t-1}^i||^2)$, which measures the average inter-frame joint moving distance, indicating the motion energy level and pose stability. The proposed mmDiff demonstrates enhanced pose stability by minimizing joint vibration and avoiding abrupt pose changes.

Table 5: Model efficiency evaluated on mmBody. We compare mmDiff's diffusion training in phase two with the benchmark method P4Transformer [12]. Extra computational resources of our designed modules for one diffusion step are illustrated.

Modules	Input	Input Size	Latency	#Params.	GFLOPs.
P4Transformer [12]	R	$(N, 6)$	40.48 ms	128.00M	43.50
Diffusion model (D)	$\hat{H}_k$	$(17, 3)$	7.59 ms	1.03M	0.03
Global context (G)	F^j	$(17, 64)$	0.36 ms	0.02M	0.01
Local context (L)	$R, \hat{H}_k$	$(N, 6), (17, 3)$	2.00 ms	0.19M	0.40
Motion consistency (T)	$\{\tilde{H}_{t-i}\}_{i=1}^{\Delta t}$	$(\Delta t, 17, 3)$	1.74 ms	15.98M	0.19
Limb consistency (S)	F^j	$(17, 64)$	0.16 ms	1.29M	0.02
(D+G+L+T+S)	/	/	11.85 ms	18.51M	0.62

Effects of Structural Limb-length Consistency. To validate the effectiveness of the limb-length consistency, we compare mmDiff w/ and w/o the module (C^{lim}) and plot the histograms that indicate the limb-length distribution in Figure 5. The ground truth limb length remains constant for all limbs. We observe both reduced limb-length error (in the estimated arms, legs, and spline) and variance (in forearms and lower legs), indicating enhanced pose accuracy and structural stability. We argue that more accurate limb-length can lead to more accurate pose estimation. Qualitatively in figure 4, the variant height and arm's length for chest expansion and hand raising are mitigated with mmDiff. However, as in Table 1, SLC has limitations when the sensing distance is long (e.g., in Lab 2) due to harder limb-length estimation.

Model Efficiency. We provide the efficiency analysis in Table 5, where our proposed modules require substantially little computational resources during phase-two diffusion training and inference. All designed modules have a latency of less than 2ms, demonstrating outstanding efficiency in supporting the iterative diffusion process. Meanwhile, the model's parameters of our designed modules are relatively small compared to the P4Transformer [12], and require minimum computation complexity, as reflected by GFLOPs. It demonstrates the potential for applications like robotics, the Internet of Things, and edge computing.

5 Conclusion

This paper proposes mmDiff as a human pose estimation (HPE) method for RF-vision, a conditional diffusion model designed to generate accurate and stable human poses from noisy mmWave radar point clouds. The proposed modules demonstrate enhanced robustness in handling radar's miss-detection and signal inconsistency when extracting the guiding features. Compared to the state-of-the-art methods, the proposed mmDiff demonstrates better accuracy and stability in mmWave HPE, showing the viability of radar-based HPE to deal with low illumination or haze. However, it also has limitations worth future studies, including slightly inferior performance compared to RGB-D and increased noise in the presence of multiple sensing targets.

Acknowledgements

This research is supported by the National Research Foundation of Singapore under its Medium-Sized Center for Adavnced Robotics Technology Innovation, and Ministry of Education of Singapore under ACRF Tier 1 Grant RG 64/23.

References

1. An, S., Li, Y., Ogras, U.: mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. Advances in Neural Information Processing Systems **35**, 27414–27426 (2022)
2. An, S., Ogras, U.Y.: Fast and scalable human pose estimation using mmwave point cloud. In: Proceedings of the 59th ACM/IEEE Design Automation Conference. pp. 889–894 (2022)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014)
4. Ar, I., Akgul, Y.S.: A computerized recognition system for the home-based physiotherapy exercises using an rgbd camera. IEEE Transactions on Neural Systems and Rehabilitation Engineering **22**(6), 1160–1171 (2014)
5. Bansal, K., Rungta, K., Zhu, S., Bharadia, D.: Pointillism: Accurate 3d bounding box estimation with multi-radars. In: Proceedings of the 18th Conference on Embedded Networked Sensor Systems. pp. 340–353 (2020)
6. Chen, A., Wang, X., Zhu, S., Li, Y., Chen, J., Ye, Q.: mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3501–3510 (2022)
7. Chen, H., Feng, R., Wu, S., Xu, H., Zhou, F., Liu, Z.: 2d human pose estimation: A survey. Multimedia Systems **29**(5), 3115–3138 (2023)
8. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Transactions on Circuits and Systems for Video Technology **32**(1), 198–209 (2021)
9. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018)
10. Deng, T., Xie, H., Wang, J., Chen, W.: Long-term visual simultaneous localization and mapping: Using a bayesian persistence filter-based global map prediction. IEEE Robotics & Automation Magazine **30**(1), 36–49 (2023)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Fan, H., Yang, Y., Kankanhalli, M.: Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14204–14213 (2021)
13. Gao, Q., Liu, J., Ju, Z., Zhang, X.: Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation. IEEE Transactions on Industrial Electronics **66**(12), 9663–9672 (2019)
14. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13041–13051 (2023)

15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
16. Hu, M., Wang, Y., Cham, T.J., Yang, J., Suganthan, P.N.: Global context with discrete diffusion in vector quantised modelling for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11502–11511 (2022)
17. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013)
18. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Advances in Neural Information Processing Systems **35**, 23593–23606 (2022)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5253–5263 (2020)
21. Lee, S.P., Kini, N.P., Peng, W.H., Ma, C.W., Hwang, J.N.: Hupr: A benchmark for human pose estimation using millimeter wave radar. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5715–5724 (2023)
22. Li, G., Zhang, Z., Yang, H., Pan, J., Chen, D., Zhang, J.: Capturing human pose using mmwave radar. In: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). pp. 1–6. IEEE (2020)
23. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
24. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2640–2649 (2017)
25. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. Acm transactions on graphics (tog) **36**(4), 1–14 (2017)
26. Prabhakara, A., Jin, T., Das, A., Bhatt, G., Kumari, L., Soltanaghai, E., Bilmes, J., Kumar, S., Rowe, A.: High resolution point clouds from mmwave radar. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4135–4142. IEEE (2023)
27. Qi, J., Du, J., Siniscalchi, S.M., Ma, X., Lee, C.H.: Analyzing upper bounds on mean absolute errors for deep neural network-based vector-to-vector regression. IEEE Transactions on Signal Processing **68**, 3411–3422 (2020)
28. Qiu, Z., Yang, Q., Wang, J., Wang, X., Xu, C., Fu, D., Yao, K., Han, J., Ding, E., Wang, J.: Learning structure-guided diffusion model for 2d human pose estimation. arXiv preprint arXiv:2306.17074 (2023)
29. Ramakrishna, V., Kanade, T., Sheikh, Y.: Tracking human pose by tracking symmetric parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3728–3735 (2013)
30. Rao, S.: Introduction to mmwave sensing: Fmcw radars. Texas Instruments (TI) mmWave Training Series pp. 1–11 (2017)

31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
32. Saadatnejad, S., Rasekh, A., Mofayezi, M., Medghalchi, Y., Rajabzadeh, S., Mordan, T., Alahi, A.: A generic diffusion-based approach for 3d human pose prediction in the wild. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8246–8253. IEEE (2023)
33. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. arXiv preprint arXiv:2303.11579 (2023)
34. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
35. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
36. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
37. Sun, Y., Huang, Z., Zhang, H., Cao, Z., Xu, D.: 3drimr: 3d reconstruction and imaging via mmwave radar based on deep learning. In: 2021 IEEE International Performance, Computing, and Communications Conference (IPCCC). pp. 1–8. IEEE (2021)
38. Tao, T., Yang, X., Xu, J., Wang, W., Zhang, S., Li, M., Xu, G.: Trajectory planning of upper limb rehabilitation robot based on human pose estimation. In: 2020 17th International Conference on Ubiquitous Robots (UR). pp. 333–338. IEEE (2020)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
40. Waldschmidt, C., Hasch, J., Menzel, W.: Automotive radar—from first efforts to future systems. IEEE Journal of Microwaves **1**(1), 135–148 (2021)
41. Wang, J., Jin, S., Liu, W., Liu, W., Qian, C., Luo, P.: When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11855–11864 (2021)
42. Xue, H., Cao, Q., Miao, C., Ju, Y., Hu, H., Zhang, A., Su, L.: Towards generalized mmwave-based human pose estimation through signal augmentation. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. pp. 1–15 (2023)
43. Xue, H., Ju, Y., Miao, C., Wang, Y., Wang, S., Zhang, A., Su, L.: mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave. In: Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services. pp. 269–282 (2021)
44. Yang, J., Huang, H., Zhou, Y., Chen, X., Xu, Y., Yuan, S., Zou, H., Lu, C.X., Xie, L.: Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. arXiv preprint arXiv:2305.10345 (2023)
45. Yang, J., Zhou, Y., Huang, H., Zou, H., Xie, L.: Metafi: Device-free pose estimation via commodity wifi for metaverse avatar simulation. In: 2022 IEEE 8th World Forum on Internet of Things (WF-IoT). pp. 1–6. IEEE (2022)
46. Zhang, J., Xi, R., He, Y., Sun, Y., Guo, X., Wang, W., Na, X., Liu, Y., Shi, Z., Gu, T.: A survey of mmwave-based human sensing: Technology, platforms and applications. IEEE Communications Surveys & Tutorials (2023)

47. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021)
48. Zhao, M., Li, T., Abu Alsheikh, M., Tian, Y., Zhao, H., Torralba, A., Katabi, D.: Through-wall human pose estimation using radio signals. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7356–7365 (2018)
49. Zhao, W., Wang, W., Tian, Y.: Graformer: Graph-oriented transformer for 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20438–20447 (2022)
50. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11656–11665 (2021)
51. Zhou, J., Zhang, T., Hayder, Z., Petersson, L., Harandi, M.: Diff3dhpe: A diffusion model for 3d human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2092–2102 (2023)
52. Zhou, Y., Huang, H., Yuan, S., Zou, H., Xie, L., Yang, J.: Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. IEEE Internet of Things Journal (2023)
53. Zhou, Y., Yang, J., Huang, H., Xie, L.: Adapose: Towards cross-site device-free human pose estimation with commodity wifi. arXiv preprint arXiv:2309.16964 (2023)
54. Zhu, Y., Li, Z., Wang, T., He, M., Yao, C.: Conditional text image generation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14235–14245 (2023)