mmDiffusion: mmWave Diffusion for Sequential 3D Human Dense Point Cloud Generation

Qian Xie
University of Leeds
Leeds, UK
q.xie2@leeds.ac.uk

Xinyu Hou
University of Oxford
Oxford, UK
xinyu.hou@cs.ox.ac.uk

Qianyi Deng
University of Oxford
Oxford, UK
qianyi.deng@cs.ox.ac.uk

Amir Patel
University College London
London, UK
amir.patel@ucl.ac.uk

Niki Trigoni
University of Oxford
Oxford, UK
niki.trigoni@cs.ox.ac.uk

Andrew Markham
University of Oxford
Oxford, UK
andrew.markham@cs.ox.ac.uk

Abstract

Millimeter-wave (mmWave) point-cloud radar shows great promise in enabling responsive human-machine interfaces (e.g., through pose and gesture tracking and for emerging augmented reality approaches). However, generating dense and temporally consistent 3D human point clouds from sequential mmWave signals is challenging due to point-cloud sparsity, jitter, and noise. Existing approaches have made progress in single-frame densification, but are inaccurate over multiple frames. This work redefines the problem as a 3D point cloud denoising task, leveraging reverse diffusion processes to transform sparse mmWave data into detailed and accurate whole-body representations. Our proposed method, mmDiffusion, effectively exploits diffusion models and temporal context within mmWave sequences to learn the denoising process, resulting in denser and temporally coherent human point clouds. For the first time, we also introduce an evaluation metric tailored to measure temporal consistency for sequential 3D human point clouds. Experimental results demonstrate that mmDiffusion significantly outperforms existing methods.

1. Introduction

The potential use-cases of millimeter wave (mmWave) point clouds in human monitoring, motion capture, and machine interaction are substantial, owing to the unique characteristics of mmWave radar such as highly accurate ranging and sensitivity to subtle movements [15, 31]. For instance, mmWave-based sensing has demonstrated promising results in vital sign monitoring [2, 32], gesture recognition [22], and human activity analysis [20, 39, 42], amongst other domains. However, a major challenge lies in the sparse nature of mmWave point clouds, which impedes the development of corresponding applications. To address this limitation and fully harness the capabilities of mmWave-based human sensing, researchers have focused on generating high-quality human point clouds from mmWave radar data [3, 41]. Some previous studies have made significant progress in this direction, employing advanced signal processing and machine learning techniques to enhance the density of mmWave point clouds, thereby enabling more accurate and detailed human representation [35].

Despite these promising results, existing methods face a crucial issue when dealing with mmWave signal *sequences*. Existing methods are typically designed and trained on individual mmWave frames independently, overlooking the continuous nature of mmWave signal acquisition. In practice, mmWave radar signals are captured as sequences of consecutive frames over time. Consequently, methods exclusively trained on isolated frames often exhibit unsatisfactory temporal stability, leading to perceptible visual artifacts such as flickering over time in the resulting dense point clouds. This temporal inconsistency poses a significant challenge in real-world applications where reliable and coherent human representations are essential for precise analysis and decision-making.

In this paper, we present a novel and effective approach to address the challenge of temporal consistency in 3D human point cloud generation from mmWave signal sequences, as in Fig. 1. To the best of our knowledge, this is the first attempt to tackle this crucial aspect explicitly. Our method consists of two main parts: the mmWave feature encoding part and the point clouds generation part. For the first part, we devise an mmWave temporal encoder, which

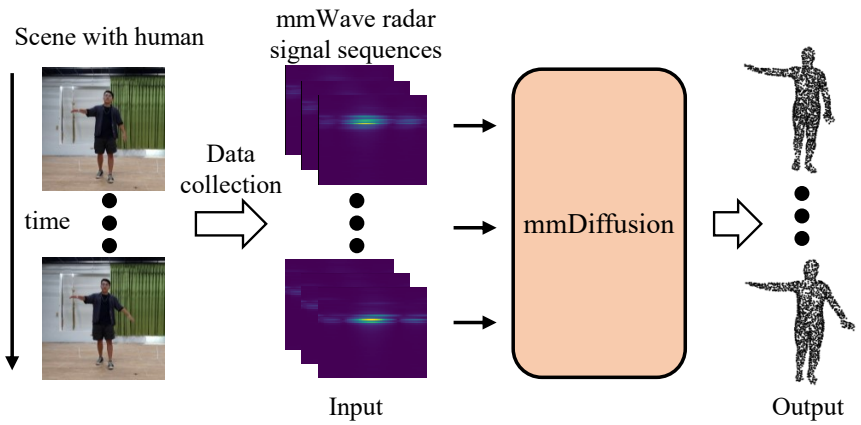

Figure 1. We design a novel architecture, called mmDiffusion, to generate temporally consistent 3D human dense point clouds from sequential mmWave signals.

combines an mmWave encoder with Gated Recurrent Units (GRUs). The mmWave encoder is responsible for learning mmWave features, while the GRUs capture the temporal dependencies present in human motion patterns within the signal sequences. By effectively integrating the two, our proposed temporal encoder enables the extraction of informative and coherent representations from the mmWave signal sequences. In the second part, we reformulate the point cloud generation task as a point cloud denoising problem. We then employ the diffusion technique to solve this point cloud denoising problem. Moreover, to realize the generation of 3D human dense point clouds, human-oriented mmWave signals are acting as conditions to guide the diffusion procedure, resulting an mmWave-conditioned diffusion model. The mmWave-conditioned diffusion model leverages the learned temporal context from our mmWave temporal encoder to resolve ambiguities present in the point cloud generation process. This results in smoother and more temporally consistent outputs, as quantified by our proposed measure of temporal consistency. By conducting extensive experiments on public benchmarks, we validate the efficacy of our approach. Our results demonstrate that the integration of temporal information significantly enhances the quality and stability of the generated dense point clouds, providing a notable advancement in the field of mmWave-based human sensing and point cloud generation. The contributions of this work are three-fold:

- For the first time, we present a novel formulation of mmWave-based 3D human point cloud generation as a 3D point cloud denoising problem. By leveraging the concept of diffusion, the proposed method tackles the sparsity issue in mmWave point clouds, effectively enhancing the density and quality of the generated point clouds.
- We introduce mmDiffusion, a novel and effective method for dense 3D human point cloud generation from sequential mmWave signals. Leveraging diffusion models, mmDiffusion efficiently learns the denoising process for point clouds, exploiting the temporal context in the sequential mmWave signals to achieve superior results.
- To evaluate the stability of sequential 3D human point cloud generation results over frames, we design a new evaluation metric. This metric provides a quantifiable measure of temporal consistency in the generated point clouds, enabling a comprehensive assessment of the performance of the proposed method in capturing coherent human motion patterns.

2. Related works

mmWave point clouds. mmWave radars, renowned for their robustness against adverse environments, cost-effectiveness, and privacy-preserving nature, have become integral in human-centric applications [33], notably in human motion sensing [8, 16, 45], gesture and activity recognition [1, 34, 38], human tracking and identification [5, 43, 44], human pose estimation [12, 13] and human mesh reconstruction [36, 37].

Operating using radio frequency (RF) signals in the tens of GHz range, these radars discern three-dimensional spatial information of detected objects, facilitating the generation of detailed 3D point clouds [6, 26]. By analyzing these point clouds, mmWave radars enable intricate understanding and intelligent sensing of targets, thus addressing both privacy concerns and the need for spatial cognition in modern applications. For instance, mmFall [11] employs an mmWave radar sensor to obtain the human body's point cloud data, which is subsequently processed using a Hybrid Variational RNN AutoEncoder (HVRAE) for human fall detection. RadHAR [29] introduces a framework for human activity recognition using sparse and non-uniform point clouds from millimeter-wave (mmWave) radars. However, low-cost mmWave radar systems produce point clouds that

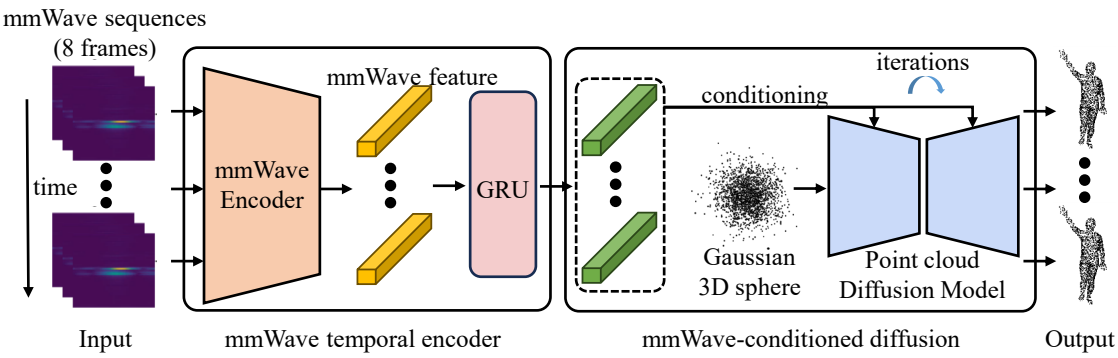

Figure 2. Architecture of the proposed mmDiffusion for 3D human dense point clouds generation from sequential mmWave signals. The network consists of two main components: an mmWave temporal encoder and an mmWave-conditioned diffusion module. The temporal encoder takes mmWave sequences as input and outputs mmWave features with temporal information. The diffusion module then generates 3D point clouds under the conditioning of these mmWave features. GRU: Gated Recurrent Unit.

are sparse (e.g. 64-128 points) and not uniformly sampled, making fine-grained activity classification challenging. In addition, there is significant background noise from reflections and small movements. Recently, mmPoint [35] focuses on generating dense human 3D point clouds from mmWave signals via formulating the point cloud generation task as a point cloud deformation problem, which is solved by a first-densify-then-deform strategy performed on a template human point cloud. However, temporal information, which is important for applications involving human motion analysis, is not taken into consideration in mmPoint, resulting in its poor performance on sequential mmWave signals. In this paper, instead of processing single-frame mmWave signals separately, we propose to introduce temporal information to ensure the consistency of 3D shapes along sequential mmWave signals.

3D diffusion models. Diffusion models, inspired by nonequilibrium thermodynamics [30], have emerged as a more promising generative modelling paradigm than GANs [9]. The pioneering work, Denoising Diffusion Probabilistic Models (DDPM) [10], introduces this class of generative models and applies them to image generation tasks, demonstrating their ability to produce high-quality image synthesis results. After that, several image generation methods based on diffusion models have been developed [7, 21, 27, 28]. Encouraged by these achievements in image generation, researchers have extended diffusion models to the realm of 3D shape generation [4, 25], within 3D diffusion model-based frameworks. This adaptation showcases the versatility and promise of diffusion models in the generation of complex 3D structures, offering new possibilities in the field of 3D content synthesis.

Focusing on 3D point clouds, [17] introduces diffusion models into the 3D point cloud generation task for the first time, by viewing points in point clouds as particles in a thermodynamic system. In [46], a novel probabilistic generative model called Point-Voxel Diffusion (PVD) is introduced, which combines denoising diffusion models with a hybrid point-voxel representation to generate high-fidelity 3D shapes and complete partial point clouds. Point Diffusion-Refinement (PDR) paradigm [18] is introduced for 3D point cloud completion, combining a Conditional Generation Network (CGNet) for coarse completion with a ReFinement Network (RFNet) for refinement. Similar to our work, [19] presents PC2, which leverages projection-conditioned point cloud diffusion to reconstruct 3D point clouds from a single RGB image, achieving high-resolution geometries closely aligned with the input image. Distinct from the aforementioned approaches that primarily focus on RGB images or traditional visual cues, our work leverages mmWave radar signals to condition a point cloud diffusion model, enabling the generation of dense point clouds specifically tailored to human targets, presenting a novel intersection of radar technology and point cloud generation.

3. Method

The overall framework of mmDiffusion is summarised in Fig. 2. Our framework comprises two key components: the mmWave temporal encoder and the mmWave-conditioned diffusion. For a single-person input mmWave signal sequence of length T frames ($T = 8$ in our experiments), we initially employ an mmWave encoder to extract frame-level features. Next, we train a temporal encoder, integrating Gated Recurrent Units (GRUs) to capture temporal dependencies, generating latent variables with past and future context. These augmented mmWave features, combining temporal information, are then used to condition a point cloud diffusion model for dense 3D point cloud generation. This comprehensive framework ensures that temporal coherence is embedded within the generated point clouds, resulting in smoother and visually consistent representations

of human motion over time.

3.1. mmWave Temporal Encoder

In the first place, we extract informative features from each frame $F_t(t = 1, \dots T)$ of the mmWave signal sequence. To achieve this, we employ the mmWave encoder in [35], which is a convolutional network and outputs a vector $f_t \in \mathbb{R}^{128}$. This encoder is adept at capturing the intricate nuances of mmWave data, enabling it to generate rich frame-level representations. More details about the architecture of this encoder can be referred to [35]. With the frame-level features $f_1, \dots f_T$ in hand, we proceed to train a specialized temporal encoder. This temporal encoder is a dynamic component, comprising a Gated Recurrent Unit (GRU) layer that yields latent feature vectors $l_1, \dots l_T \in \mathbb{R}^{128}$. The GRU layer plays a pivotal role in capturing the temporal dependencies present in the sequential mmWave signals. Importantly, the bidirectional architecture ensures that information from both past and future frames is incorporated into the encoding process. The output of this temporal encoder comprises latent variables that embody the evolving dynamics of human motion throughout the sequence. The output from our mmWave temporal encoder combines the inherent mmWave features with the crucial temporal context captured by the bidirectional GRUs. This enriched representation of mmWave data is instrumental in overcoming the sparsity challenge associated with mmWave point clouds, as it encapsulates both spatial and temporal information.

3.2. mmWave-Conditioned Diffusion

Armed with the augmented mmWave features, we proceed to generate dense 3D point clouds. To achieve this, we employ a point cloud diffusion model, which is conditioned on the mmWave features enriched with temporal information.

Diffusion models unconditionally generate meaningful content via learning a Markov Chain that gradually transfers a simple distribution, such as isotropic Gaussian $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, into a data distribution [10]. Given a 3D point cloud with N points $P_0 \in R^{3\times N}$ sampled from a data distribution $q(P_0)$ representing a human with posture, a 3D Gaussian noise point cloud P_S can be obtained by gradually adding noise to P_0 with S steps, forming a series of latent variables $P_1, \dots, P_S$ of the same dimensionality. At each step, the noise addition (i.e., diffusion) process can be seen as a Gaussian translation:

$$q\left(P_s \mid P_{s-1}\right) = \mathcal{N}\left(\sqrt{1-\beta_s}P_{s-1}, \beta_s \boldsymbol{I}\right), \tag{1}$$

where $\{\beta_s\}_{s=1}^{T}$ is a fixed variance schedule. Given Eq. 1, sampling P_s from P_0 for any timestep s can be written as:

$$q\left(P_s \mid P_0\right) = \mathcal{N}\left(\sqrt{\bar{\alpha}_s}P_0, \left(1-\bar{\alpha}_s\right)\boldsymbol{I}\right), \tag{2}$$

where $\alpha_s := 1 - \beta_s$ and $\bar{\alpha}_s := \prod_{s=1}^{s} \alpha_s$. From Eq. 2, we then have:

$$P_s = \sqrt{\bar{\alpha}_s}P_0 + \sqrt{1-\bar{\alpha}_s}\epsilon, \tag{3}$$

where $\epsilon \sim N(0, \mathbf{I})$ has the same dimensionality as data P_0. The reverse $q\left(P_{s-1} \mid P_s\right)$ of the above diffusion process can then be used as a generative model via denoising an arbitrary Gaussian noise to a data distribution. Since $q\left(P_{s-1} \mid P_s\right)$ is intractable, we then train a network f_θ to fit the reversal process, which is equivalent to learning to predict the noise ϵ in Eq. 3. Given a clean sample input P_0 and a random Gaussian noise ϵ, the noise prediction network f_θ can be learned by minimizing the L_2 distance between the true noise and the predicted noise:

$$L(\theta) = \mathbb{E}_{s, P_0, \epsilon}\left[\left\|\epsilon - f_\theta\left(\sqrt{\bar{\alpha}_s}P_0 + \sqrt{1-\bar{\alpha}_s}\epsilon, s\right)\right\|^2\right] \tag{4}$$

Once the network is trained well, we can then sample a noise latent variable P_{S-1} from the distribution $q\left(P_{s-1} \mid P_s\right)$. Iteratively, we can finally arrive at a clean point cloud P_0. However, this denoising process is unconditional [10], which means we cannot control the shape of the final generated point clouds.

In this paper, we formulate the 3D human point cloud reconstruction task as a conditional generation problem. That is, instead of learning the original vanilla target distribution $q(P_0)$, we propose to learn a conditional distribution $q(P_0|M)$ where M represents an mmWave signal by the noise prediction network $f_\theta(P_0, s, M)$. Following [46], a Point-Voxel CNN (PVCNN) [14] is adopted as the base of our noise prediction network f_θ. The architecture of f_θ is illustrated in Fig. 3. As shown, we have four PVConv-SA layers to downsample points and extract features and four PVConv-FP layers to upsample points and propagate features. For step s, the point cloud diffusion network takes noisy point clouds P_s as input, and predicts the noise. We can then get a point cloud P_{s-1} with less noise for step $s-1$.

To condition the denoising (i.e., the noise prediction process), we design an mmWave Conditioning (MMC) layer to manipulate point features with mmWave features. MMC layers are placed after each SA and FP layer. The architecture of MMC is illustrated in Fig. 3. Given the intermediate point feature $f_p \in R^{n_f \times n_p}$ produced by SA or FP layers, as input, MMC aims to perform a feature-wise affine transformation on it to get a conditioned point feature $f_p^{'}$. n_f and n_p are the number of feature channels and the number of points. Inspired by the general conditioning layer proposed in [24], MMC learns two fully connected layers which take the mmWave feature $f_m \in R^{128}$ as input and output conditioning parameters $\lambda \in R^{n_f}$ and $\beta \in R^{n_f}$. We then substitute the conditioned version feature $f_p^{'}$ for f_p:

$$f_p^{'} = \mathrm{ReLU}\left[(1+\gamma) \odot f_p \oplus \beta\right], \tag{5}$$

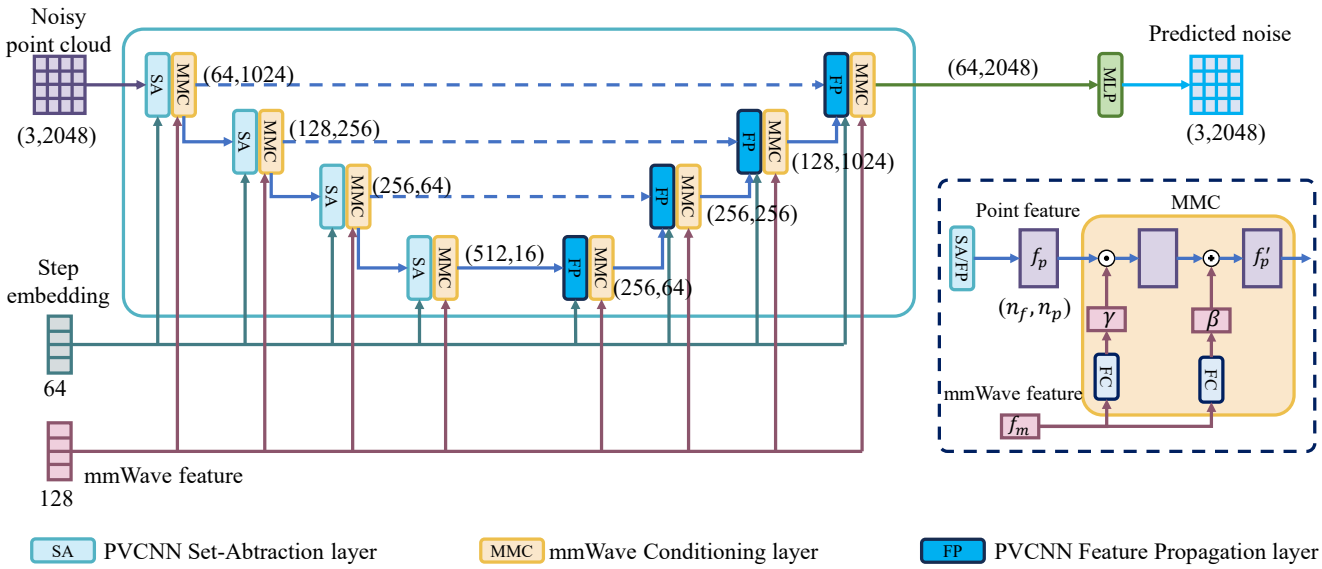

Figure 3. Architecture of the proposed mmWave-conditioned diffusion model for point cloud generation. We employ a Point-Voxel CNN (PVCNN) [14] as the base encoder and design an mmWave conditioning (MMC) layer to integrate mmWave information to the point cloud generation procedure. SA and FP layers are native components within the PVCNN architecture. More details about this two layers can be referred to [14].

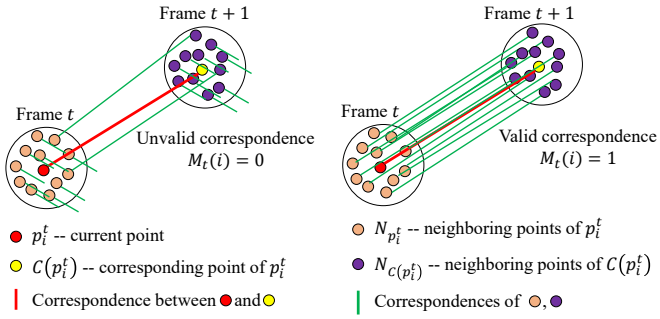

Figure 4. Illustration of how the valid correspondence mask M_t is established. When the correspondence of a point p_i^t is not consistent with the correspondences of the point's neighboring points, the correspondence is a wrong correspondence, i.e., $M_t(i) = 0$. On the contrary, it is a valid correspondence and $M_t(i) = 1$.

where $\odot$, $\oplus$ represent the elementwise multiplication and addition operations. In this way, the original point feature f_p is scaled by λ and translated by β.

3.3. Metric of Temporal Consistency

To assess the stability of sequential point cloud data, it is crucial to quantify the variations in 3D space across the entire point cloud sequence. Thus, we present a metric of stability measurement for human point cloud sequences.

We first employ a human-oriented point cloud correspondence method [40] to establish a correspondence map, denoted as C, between two consecutive point clouds, P_t and P_{t+1}. With this correspondence map, we can readily identify corresponding points between the two frames. Specifically, for each point p_i^t in P_t, its corresponding point in P_{t+1} can be represented as $C(p_i^t)$. However, we notice that the correspondence map C established by [40] may not be perfect. Thus, it is necessary to identify and remove these wrong correspondences to obtain a valid mask, M_t, indicating which point has a correct correspondence. To construct a valid mask M_t, we implement a correspondence score by comparing the correspondences of each point p_i^t with those of its k-nearest neighbors, denoted as $N_{p_i^t}$, as illustrated in Fig. 4. k is 20 in our paper. Specifically, the validity of the correspondence for p_i^t is assessed by ensuring consistency between $C(p_i^t)$ and the correspondences $C(N_{p_i^t})$ of its neighboring points $N_{p_i^t}$. Let p_{ij}^t be the j-th point in $N_{p_i^t}$, we introduce a binary indicator function, denoted as $I(p_{ij}^t)$, to determine whether the corresponding point of a neighboring point p_{ij}^t lies within the neighborhood of $N_{C(p_i^t)}$. Formally, $I(p_{ij}^t)$ equals 1 if $C(p_{ij}^t)$ falls within $N_{C(p_i^t)}$, and 0 otherwise. That is,

$$I(p_{ij}^t) = \begin{cases} 1, & C(p_{ij}^t) \in N_{C(p_i^t)} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

To obtain an overall correspondence score $S_c(p_i^t)$ for a point p_i^t, we sum the values of $I(p_{ij}^t)$ across all neighboring points $N_{p_i^t}$. That is,

$$S_c(p_i^t) = \sum_{j=1}^{k} I(p_{ij}^t). \tag{7}$$

Ideally, this score should be close to k if the neighboring correspondences are correct, as illustrated in the lower case in Fig. 4. In practice, when the score is below a predefined threshold thr_c (10 in our experiments), we consider the correspondence of point p_i^t as invalid.

With the valid mask M_t in place, we can now quantitatively measure the variation between point clouds P_t and P_{t+1}. By calculating the pairwise distances between corresponding points in P_t and P_{t+1}, we can obtain a vanilla measurement of the 3D spatial fluctuations across the point cloud sequence. However, this vanilla measurement based on absolute distance may fail to account for the influence of object motion speed. When a person moves slowly, even small distance variations between adjacent frames can lead to significant perceptual flickering. Conversely, when a person moves rapidly, the distance variations may be large, even if the point clouds remain consistent. To address this issue, we propose a metric that takes into consideration the ratio of change, instead of the absolute distance, thereby introducing a measure of relative difference. Specifically, given a predefined distance threshold thr_d, we calculate the percentage of points whose distance variations fall below this threshold among all valid matching points. This metric quantifies the proportion of matching points that exhibit modest variation at frame t, considering the relative changes in their distances:

$$\text{TC}_t = \frac{\sum_{i=1}^{|P_t|} M_t(i) \cdot (\|p_i^t - C(p_i^t)\|_2 < thr_d)}{\sum_{i=1}^{|P_t|} M_t(i)}. \quad (8)$$

This metric provides a more nuanced temporal consistency assessment, accounting for object motion speed and information on the percentage of matching points exhibiting modest variation in the frame t. An example is given in Fig. 11 in *supplementary material*. In summary, our proposed metric ensures robust point cloud correspondence and enables the assessment of stability by validating correspondences and measuring variations in 3D space. This methodology provides a foundation for analyzing the stability of sequential point clouds generated from mmWave signal data.

4. Experiments

4.1. Training and Dataset

PyTorch [23] is used in our paper. The model is trained with batch size of 8 for a total of 200 steps. For mmWave feature extraction, we use the mmWave encoder in mmPoint [35]. We employ a two-stage training strategy. Specifically, we first train a network without GRU. We then train the proposed network while loading and fixing the weights from the first stage training. That is, we only need to train the GRU network at the second stage. We use point clouds with 2048 points to maintain consistency with mmPoint. We will make our code publicly available upon acceptence.

Metrics	Methods	Scenes				
		#1	#2	#3	#4	#5
$CD_{L1}\downarrow$	mmMesh [36]	11.56	10.63	8.81	13.11	9.87
	mmPoint [35]	3.21	**2.87**	2.97	3.55	2.92
	ours	**3.15**	3.09	**2.78**	**3.27**	**2.73**
TC↑	mmMesh [36]	-	-	-	-	-
	mmPoint [35]	0.615	0.609	0.535	0.517	0.560
	ours	**0.774**	**0.808**	**0.689**	**0.721**	**0.739**

Table 1. Comparison of 3D dense human point cloud generation performance.

We use the dataset in [35] which proposes to generate pesudo-ground truth point clouds by image-based 3D human mesh reconstruction on the HuPR [13] dataset. The original dataset in mmPoint contains human mmWave signal-point cloud pairs for 58 scenes, each of which consists of 600 frames data. Given that our proposed network possesses a higher parameter count than mmPoint, we enhance the original dataset for a more comprehensive training. Specifically, we augmented the dataset to 70 scenes using the dataset generation approach in mmPoint. Among them, 5 scenes are used for testing. We employ the L1 version of the Chamfer distance$\times 10^2$ (CD_{L1}) and our proposed temporal consistency (TC) as quantitative evaluation metrics to assess the efficacy of the proposed methodology. For TC, the distance threshold thr_d is 0.3, and we compute the average value for 599 adjacent frame pairs in the whole 600 frames for each scene.

4.2. Comparison.

Tab. 1 compares the performance of the proposed novel method for 3D dense human point cloud generation from mmWave signals against the most related prior work, mmPoint [35]. We also compare with mmMesh [36], in which a traditional method is designed to generate 3D point cloud from mmWave signals and a point-to-mesh network is proposed to finally get 3D human mesh from the 3D point clouds. Following mmPoint, we focus on the 3D point cloud generation part in mmMesh since it is more related to our method. The results are segmented into five different scenes for a comprehensive analysis. The proposed method demonstrates an improvement over mmPoint in all scenes for both metrics, except for scene #2. For CD_{L1}, the improvements range from 0.08 to 0.28 across the scenes, with our method consistently yielding lower distances, indicating better accuracy of the point cloud reconstruction. In terms of TC, mmDiffusion shows a significant increase in percentage points across all scenes, with improvements ranging from 11.4% to 17.9%, demonstrating a more consistent moving action of the sequential point clouds. These results suggest that the proposed method is capable of generating more accurate and more consistent human point clouds from sequential mmWave signals compared to the existing

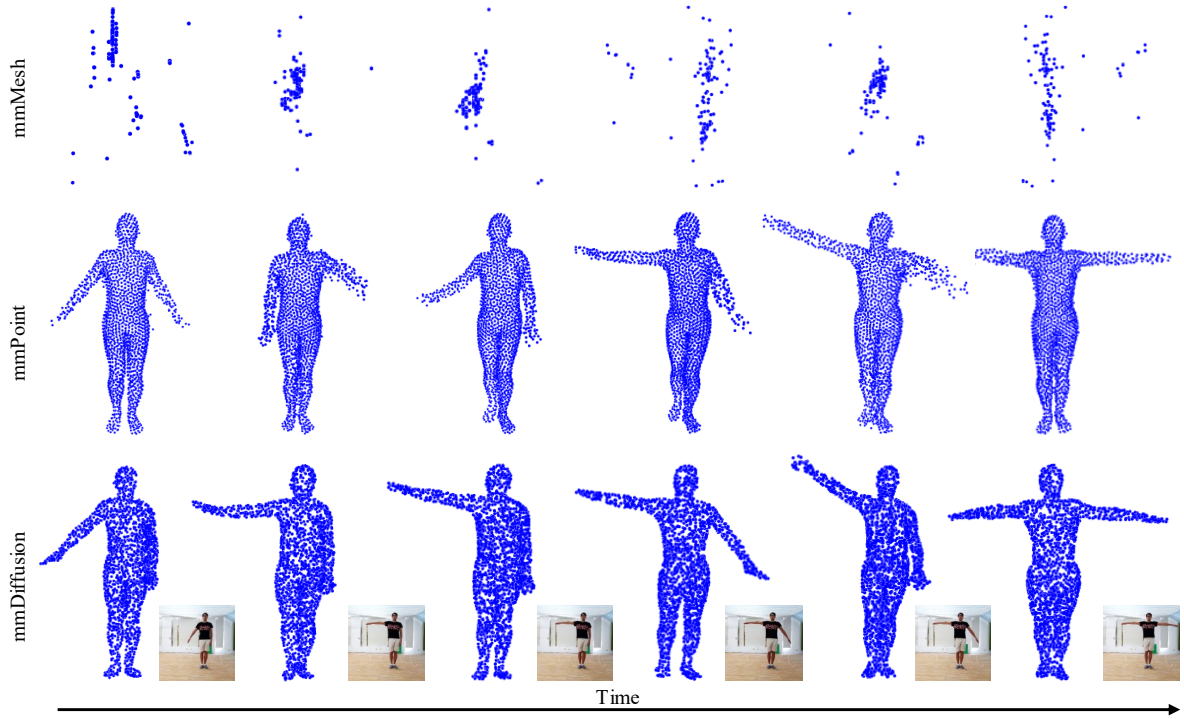

Figure 5. Visual comparison on 3D human point clouds between different methods.

approach. Fig. 5 qualitatively shows the comparison of the previous state-of-the-art mmPoint and our method.

4.3. Ablation study

GRU. In our investigation into the significance of temporal information, we employed a GRU to encode these temporal dynamics. The results, as presented in Tab. 2a, emphatically demonstrate the benefits of this approach. Without GRU, the generation results are 3.18, and 0.658 for metrics CD_{L1} and TC. However, when equipped with GRU to encode temporal information, the performance improved markedly. This enhancement underscores the pivotal role of temporal dynamics in mmWave features, and the efficacy of the GRU in capturing and leveraging these dynamics for more accurate and robust point cloud generation. We also give visual results with and without GRU in Fig. 6. As shown, integrating temporal information helps improve the stability of the output results.

Conditioning strategy. We assess the efficacy of our MMC layer, in comparison to two baseline strategies: concat-V1 and concat-V2. The concat-V1 strategy is a straightforward method wherein the mmWave feature and the input point cloud are directly concatenated before they are fed into the point cloud diffusion model. Concat-V2 follows the same overall architecture as ours, but within its conditional layers, mmWave features are simply fused via direct concatenation with point features. *Detailed architectures is in Sec. 6 of the supplementary material.* As in Tab. 2b, our MMC significantly outperforms the baseline methods. For CD_{L1}, MMC achieves 3.00, markedly better than the 3.13 and 3.07 produced by concat-V1 and concat-V2. A similar trend is observed in the TC metric. The superior performance of MMC can be attributed to its intricate design that enables more effective integration of mmWave features with point cloud data. Instead of mere concatenation, MMC leverages a mechanism that possibly allows for adaptive weighting and transformation of the point cloud

Metrics	$CD_{L1}\downarrow$	TC↑
w/o GRU	3.18	0.658
w/ GRU	3.00	0.746

(a)

Condition strategy	$CD_{L1}\downarrow$	TC↑
concat-V1	3.13	0.702
concat-V2	3.07	0.718
MMC (Ours)	3.00	0.746

(b)

Table 2. Ablation study results. (a) Impact of encoding temporal information using a GRU on the mmWave features. (b) Performance of the conditioning layers. All the results are average values over 5 testing scenes.

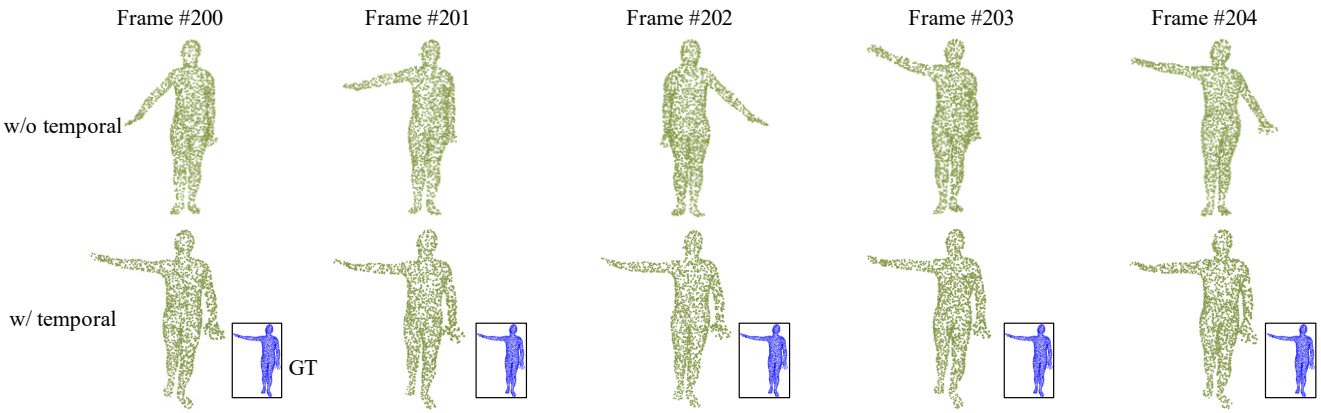

Figure 6. Comparison results with and without GRU (temporal). For the given five adjacent frames, our method with temporal binding produces more consistent results.

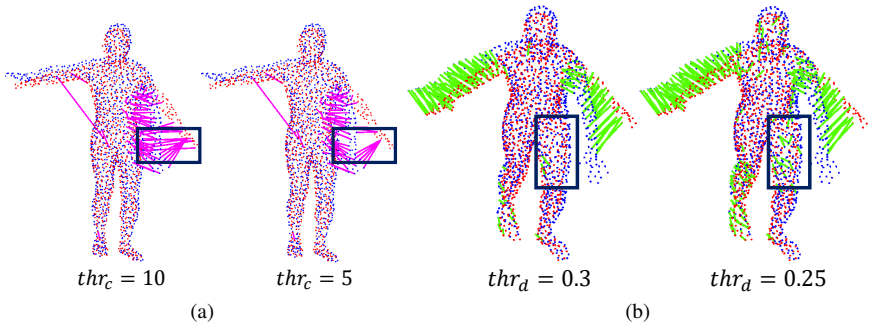

Figure 7. (a) A smaller thr_c could lead to missing some wrong correspondences. (b) A smaller thr_d is more inclined to include more noisy correspondences.

features based on the mmWave information, leading to a more accurate point cloud diffusion outcome.

Thresholds in TC. There are two thresholds in TC: the correspondence score threshold thr_c and the distance threshold thr_d. The purpose of thr_c is to eliminate wrong correspondences. Small values may lead to the retention of surplus wrong correspondences, as illustrated in Fig. 7a. The second parameter thr_d represents a distance tolerance, implying that if the movement of a certain body part between adjacent frames exceeds this tolerance, it will be considered as an unreasonable (i.e., inconsistent) movement distance, as in Fig. 7b. This assumption stems from the notion that, ideally, the movement distances of various body parts would be minimal over short time intervals. As elucidated above, thr_d needs to be adjusted according to different datasets. For instance, in datasets where human movements are rapid, compared to our current dataset, the movement distances of various body parts over the same time interval would also increase. This implies that the occurrence of relatively large distances in correspondences between adjacent point clouds is reasonable. Consequently, in such scenarios, it is necessary to set a larger distance threshold to prevent these relatively large correspondences from being deemed as inconsistent.

5. Conclusion

In this paper, we present a pioneering approach to address the challenges associated with 3D human dense point clouds generation from sequential mmWave signals. Firstly, we redefine the problem of 3D human point clouds generation from sequential mmWave signals as an mmWave-conditioned 3D human point clouds denoising task, bridging the gap between sparse mmWave data and coherent human representations through reverse diffusion processes. Secondly, our proposed method, mmDiffusion, leverages GRUs within the mmWave temporal encoder to capture and incorporate rich temporal dependencies, leading to dense and temporally coherent point clouds. Furthermore, our introduction of a novel evaluation metric tailored to measure temporal consistency adds a unique dimension to the assessment of generated point clouds. Experimental results underscore the exceptional performance of our framework.

References

[1] Anum Ali, Priyabrata Parida, Vutha Va, Saifeng Ni, Khuong Nhat Nguyen, Boon Loong Ng, and Jianzhong Charlie Zhang. End-to-end dynamic gesture recognition using mmwave radar. *IEEE Access*, 10: 88692–88706, 2022. 2

[2] Mostafa Alizadeh, George Shaker, João Carlos Martins De Almeida, Plinio Pelegrini Morita, and Safeddin Safavi-Naeini. Remote monitoring of human vital signs using mm-wave fmcw radar. *IEEE Access*, 7:54958–54968, 2019. 1

[3] Sizhe An and Umit Y Ogras. Fast and scalable human pose estimation using mmwave point cloud. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, pages 889–894, 2022. 1

[4] Gene Chou, Yuval Bahat, and Felix Heide. Diffusionsdf: Conditional generative modeling of signed distance functions. *arXiv preprint arXiv:2211.13757*, 2022. 3

[5] Han Cui and Naim Dahnoun. High precision human detection and tracking using millimeter-wave radars. *IEEE Aerospace and Electronic Systems Magazine*, 36(1):22–32, 2021. 2

[6] Han Cui, Shu Zhong, Jiacheng Wu, Zichao Shen, Naim Dahnoun, and Yiren Zhao. Milipoint: A point cloud dataset for mmwave radar. *arXiv preprint arXiv:2309.13425*, 2023. 2

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[8] Fangqiang Ding, Zhen Luo, Peijun Zhao, and Chris Xiaoxuan Lu. milliflow: Scene flow estimation on mmwave radar point cloud for human motion sensing. *arXiv preprint arXiv:2306.17010*, 2023. 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[11] Feng Jin, Arindam Sengupta, and Siyang Cao. mmfall: Fall detection using 4-d mmwave radar and a hybrid variational rnn autoencoder. *IEEE Transactions on Automation Science and Engineering*, 19(2):1245–1257, 2020. 2

[12] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. m3track: mmwave-based multi-user 3d posture tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pages 491–503, 2022. 2

[13] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5715–5724, 2023. 2, 6

[14] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems*, 32, 2019. 4, 5

[15] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. See through smoke: robust indoor mapping with low-cost mmwave radar. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, pages 14–27, 2020. 1

[16] Chris Xiaoxuan Lu, Muhamad Risqi U Saputra, Peijun Zhao, Yasin Almalioglu, Pedro PB De Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. milliego: single-chip mmwave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 109–122, 2020. 2

[17] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 3

[18] Zhaoyang Lyu, Zhifeng Kong, Xudong Xu, Liang Pan, and Dahua Lin. A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530*, 2021. 3

[19] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12923–12932, 2023. 3

[20] Zhen Meng, Song Fu, Jie Yan, Hongyuan Liang, Anfu Zhou, Shilin Zhu, Huadong Ma, Jianhua Liu, and Ning Yang. Gait recognition for co-existing multiple people using millimeter wave sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 849–856, 2020. 1

[21] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[22] Sameera Palipana, Dariush Salami, Luis A Leiva, and Stephan Sigg. Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(1):1–27, 2021. 1

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[24] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[25] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[26] Kun Qian, Zhaoyuan He, and Xinyu Zhang. 3d point cloud generation with millimeter-wave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–23, 2020. 2

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[29] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, pages 51–56, 2019. 2

[30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[31] Bram van Berlo, Amany Elkelany, Tanir Ozcelebi, and Nirvana Meratnia. Millimeter wave sensing: A review of application pipelines and building blocks. *IEEE Sensors Journal*, 21(9):10332–10368, 2021. 1

[32] Fengyu Wang, Feng Zhang, Chenshu Wu, Beibei Wang, and KJ Ray Liu. Vimo: Multiperson vital sign monitoring using commodity millimeter-wave radio. *IEEE Internet of Things Journal*, 8(3):1294–1307, 2020. 1

[33] Shuai Wang, Dongjiang Cao, Ruofeng Liu, Wenchao Jiang, Tianshun Yao, and Chris Xiaoxuan Lu. Human parsing with joint learning for dynamic mmwave radar point cloud. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(1):1–22, 2023. 2

[34] Yuheng Wang, Haipeng Liu, Kening Cui, Anfu Zhou, Wensheng Li, and Huadong Ma. m-activity: Accurate and real-time human activity recognition via millimeter wave radar. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8298–8302. IEEE, 2021. 2

[35] Qian Xie, Qianyi Deng, Ta-Ying Cheng, Peijun Zhao, Amir Patel, Niki Trigoni, and Andrew Markham. mmpoint: Dense human point cloud generation from mmwave. *British Machine Vision Conference (BMVC)*, 2023. 1, 3, 4, 6

[36] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 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*, pages 269–282, 2021. 2, 6

[37] Hongfei Xue, Qiming Cao, Yan Ju, Haochen Hu, Haoyu Wang, Aidong Zhang, and Lu Su. M4esh: mmwave-based 3d human mesh construction for multiple subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 391–406, 2022. 2

[38] Xiaopeng Yang, Weicheng Gao, Xiaodong Qu, Peng Yin, Haoyu Meng, and Aly E Fathy. A lightweight multi-scale neural network for indoor human activity recognition based on macro and micro-doppler features. *IEEE Internet of Things Journal*, 2023. 2

[39] Chengxi Yu, Zhezhuang Xu, Kun Yan, Ying-Ren Chien, Shih-Hau Fang, and Hsiao-Chun Wu. Noninvasive human activity recognition using millimeter-wave radar. *IEEE Systems Journal*, 16(2):3036–3047, 2022. 1

[40] Yiming Zeng, Yue Qian, Zhiyu Zhu, Junhui Hou, Hui Yuan, and Ying He. Corrnet3d: Unsupervised end-to-end learning of dense correspondence for 3d point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[41] Guangcheng Zhang, Xiaoyi Geng, and Yueh-Jaw Lin. Comprehensive mpoint: A method for 3d point cloud generation of human bodies utilizing fmcw mimo mm-wave radar. *Sensors*, 21(19):6455, 2021. 1

[42] Guangcheng Zhang, Shenchen Li, Kai Zhang, and Yueh-Jaw Lin. Machine learning-based human posture identification from point cloud data acquisitioned by fmcw millimetre-wave radar. *Sensors*, 23(16):7208, 2023. 1

[43] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. mid: Tracking and identifying people with millimeter wave radar. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 33–40. IEEE, 2019. 2

[44] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. Human tracking and identification through a millimeter wave radar. *Ad Hoc Networks*, 116:102475, 2021. 2

[45] Peijun Zhao, Chris Xiaoxuan Lu, Bing Wang, Niki Trigoni, and Andrew Markham. Cubelearn: End-to-end learning for human motion recognition from raw mmwave radar signals. *IEEE Internet of Things Journal*, 2023. 2

[46] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 3, 4