mmMesh: Towards 3D Real-Time Dynamic Human Mesh Construction Using Millimeter-Wave

Hongfei Xue[1], Yan Ju[1], Chenglin Miao[2], Yijiang Wang[1], Shiyang Wang[1], Aidong Zhang[3], Lu Su[4*]
[1] State University of New York at Buffalo, Buffalo, NY USA [2] University of Georgia, Athens, GA USA
[3] University of Virginia, Charlottesville, VA USA [4] Purdue University, West Lafayette, IN USA
Email: [1] {hongfeix, yanju, yijiangw, shiyangw}@buffalo.edu, [2] cmiao@uga.edu, [3] aidong@virginia.edu, [4] lusu@purdue.edu

ABSTRACT

In this paper, we present mmMesh, the first real-time 3D human mesh estimation system using commercial portable millimeter-wave devices. mmMesh is built upon a novel deep learning framework that can dynamically locate the moving subject and capture his/her body shape and pose by analyzing the 3D point cloud generated from the mmWave signals that bounce off the human body. The proposed deep learning framework addresses a series of challenges. First, it encodes a 3D human body model, which enables mmMesh to estimate complex and realistic-looking 3D human meshes from sparse point clouds. Second, it can accurately align the 3D points with their corresponding body segments despite the influence of ambient points as well as the error-prone nature and the multi-path effect of the RF signals. Third, the proposed model can infer missing body parts from the information of the previous frames. Our evaluation results on a commercial mmWave sensing testbed show that our mmMesh system can accurately localize the vertices on the human mesh with an average error of 2.47 cm. The superior experimental results demonstrate the effectiveness of our proposed human mesh construction system.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computer systems organization** → **Embedded and cyber-physical systems**.

KEYWORDS

Deep Learning, Human Sensing, Human Mesh Estimation, Point Cloud, Millimeter Wave

ACM Reference Format:
Hongfei Xue[1], Yan Ju[1], Chenglin Miao[2], Yijiang Wang[1], Shiyang Wang[1], Aidong Zhang[3], Lu Su[4]. 2021. mmMesh: Towards 3D Real-Time Dynamic Human Mesh Construction Using Millimeter-Wave. In *The 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '21), June 24–July 2, 2021, Virtual, WI, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3458864.3467679

Figure 1: Our mmMesh system

1 INTRODUCTION

Recently, researchers have put significant efforts towards building intelligent wireless sensing systems, which aim to perceive and understand human activities by leveraging pervasive wireless signals. Thus far, the most remarkable achievement in this effort is the construction of human skeletons from the signals reflected off the human body [16, 34, 35, 47, 49]. Having the skeletal representations, a follow-up question arises: Is the information contained in the RF signal rich enough to further reconstruct the shape of the human body from which we can tell not only the height but also the somatotype, weight, and even the gender of the monitored subject?

A recent pioneer study [48] offers a preliminary answer to the above question. In that work, the authors successfully construct the human mesh by utilizing RF signals. It is revealed that RF signals contain sufficient information for the estimation of not only the pose but also the shape of human body. By overcoming the technical challenges faced by traditional camera-based human perception solutions, such as occlusion, poor lighting, clothing, as well as privacy issues, wireless human sensing technique demonstrates the potential to enable a new generation of applications capable of supporting more sophisticated interactions between humans and their physical surroundings. Despite the inspiring findings presented in [48], the application scope of their system is limited by both hardware (i.e., a carefully assembled and synchronized bulky T-shaped antenna array [2]) and model (i.e., the model only works when there is a power distribution heatmap in 3D space which

*Lu Su is the corresponding author.

MobiSys '21, June 24–July 2, 2021, Virtual, WI, USA

ACM ISBN 978-1-4503-8443-8/21/06…$15.00
https://doi.org/10.1145/3458864.3467679

is hard to obtain; and this design prohibits itself from real-time implementation).

To tackle this problem, we propose to make use of the point cloud generated from commercial portable millimeter-wave devices, and construct dynamic 3D human mesh in real-time. Such system could facilitate a wide spectrum of real-world applications. For example, the proposed system can enable more realistic augmented reality (AR) and virtual reality (VR) applications by capturing players' real-time body shape and pose. It can also be used by law enforcement officers to assess the activity, somatotype, height, weight and gender of the criminal suspect without exposing themselves by leveraging the ability of RF signals to traverse walls.

However, to unleash the power of the information carried by mmWave signals, we have to address the following challenges. First, due to the limited numbers of antennas on the commercial mmWave radar, the generated point cloud in each frame is too sparse to accurately estimate such a complex 3D human mesh. Each frame only contains hundreds of points, among which only dozens of points are correlated to the human body. It is technically impossible to directly estimate the locations of thousands of human mesh vertices from such a sparse point cloud. Second, how to correctly associate each 3D point in the point cloud with the corresponding body segment is also very challenging. Since the points from the ambient can be mistakenly regarded as the points from the subject, and the obtained point locations can be inaccurate due to both the error-prone nature and the multi-path effect of the RF signals. Third, in some frames, the points related to a specific body segment may be absent due to the specularity [47] of the RF signal reflection. How to correctly infer these missing body segments remains a challenge.

To address the above challenges, we propose a deep learning framework, named **mmMesh**, to construct the dynamic 3D human mesh from the mmWave signals. First, mmMesh encodes a 3D human body model, which allows us to use only 86 parameters to represent a whole human mesh. The incorporation of such a human body model makes it possible to use dozens of points to infer a complex human mesh. Second, the proposed mmMesh model can dynamically locate the moving subject and focus on the points near the subject other than the points from the ambient objects. Additionally, though the information in each single point can be inaccurate, mmMesh is capable of capturing the spatial relationships among the 3D points and aligning them with their corresponding body segments. What's more, our model can discriminatively treat the points and automatically assign larger weights to the points carrying information of higher quality. Third, the proposed mmMesh model employs a recurrent neural network to infer the missing body parts from the information of the previous frames.

In order to evaluate the proposed mmMesh framework, we implement a prototype of our mmMesh system using COTS millimeter wave devices. The evaluation results show that our mmMesh system can accurately localize the vertices on the human mesh with an average error of 2.47 cm. The superior experimental results demonstrate the effectiveness of our proposed human mesh construction system. Figure 1 illustrates our proposed mmMesh system[1].

[1]Project Website: https://havocfixer.github.io/mmMesh/

2 PRELIMINARY

mmWave Radar based Point Cloud Generation: In this paper, we need to calculate the point cloud and the related properties (range, velocity, and energy) of the points from the mmWave signals to feed into the designed mmMesh model. The first step is to measure the distances between mmWave radar and the objects. As we know, mmWave radar transmits FMCW (Frequency Modulated Continuous Wave) based chirp signals, which can be characterized by a start frequency f_c, bandwidth B, and duration T_c [29]. The IF (Intermediate Frequency) signals are obtained by mixing the transmitted signals and received signals. Then, FFT operation (Range-FFT) can be performed on IF signal to separate different frequency components and thus get the distance between each object and the radar denoted as $R = \frac{cfT_c}{2B}$, where c is the speed of light and f is the frequency of IF signal. The second step is to calculate the velocities of the objects. Another FFT operation (Doppler FFT) is conducted to measure the phase changes of IF signal. Then the velocity can be calculated by $v = \frac{\lambda\omega}{4\pi T_c}$, where λ is the wavelength of the chirp signal and ω is the measured phase change between two chirps with interval of T_c. The last step is to calculate the coordinates of the points. In order to generate the coordinates (x, y, z) of the object O, angle estimation is also required after calculating the distance and velocity of the object. The angle of elevation φ and azimuth θ of the object can be calculated as: $\varphi = \sin^{-1}(\frac{\omega_z}{\pi})$ and $\theta = \sin^{-1}(\frac{\omega_x}{\cos(\varphi)\pi})$, where ω_z is the phase difference between azimuth antenna and corresponding elevation antenna after Doppler-FFT, and ω_x is the phase difference between consecutive receiving azimuth antennas after Doppler-FFT. Based on above results of R, φ and θ, the position of the object O (i.e., (x, y, z)) can be calculated as $x = R\cos(\varphi)\sin(\theta)$, $z = R\sin(\varphi)$ and $y = \sqrt{R^2 - x^2 - z^2}$.

PointNet: In our proposed deep learning model, we adopt PointNet [27] as our backbone network to extract point features from the point cloud. PointNet [27] is a pioneer work to tackle the point cloud data using deep learning method. In PointNet, multi-layer perceptrons (MLP) is leveraged to extract high-level representations from point cloud features. And max-pooling operation is applied to aggregate the representations of all the points in the point cloud.

SMPL (Parametric Human Model): To get realistic human mesh output, in our model design, we encode a 3D human body model as one of our model components. Skinned Multi-Person Linear model (SMPL) [24] is a widely used parametric human model that estimates 3D human mesh by factoring human body into shape and pose parameters. Shape parameters $\vec{\beta} \in \mathbb{R}^{10}$ can be utilized to control how individuals vary in height, weight, body proportions, etc. Pose parameters $\vec{\theta} \in \mathbb{R}^{72}$ is used for the 3D surface deforms with articulation, which can be represented by 1 global 3D rotation vector of the human mesh and relative 3D rotation of 23 joints. The output of SMPL is a triangulated mesh with 6890 vertices, which is obtained by shaping the template body vertices conditioned on $\vec{\beta}$ and $\vec{\theta}$, then articulating the bones according to the joint rotations $\vec{\theta}$ via forward kinematics, and finally deforming the surface with linear blend skinning. The key advantage of SMPL model is that it can output the locations of 6890 human mesh vertices by taking 10 shape parameters and 72 pose parameter as input.

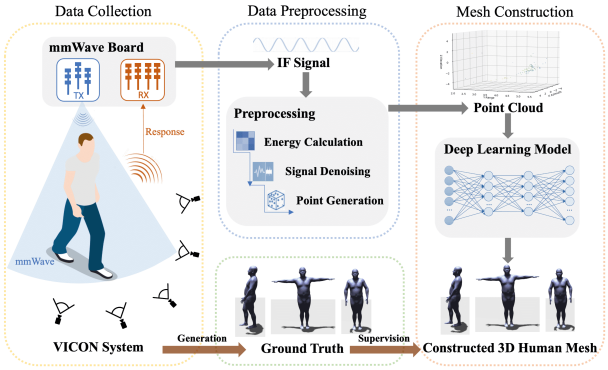

Figure 2: System Overview

3 SYSTEM OVERVIEW

In this paper, we consider a real-life scenario where the human subject is monitored by a mobile phone equipped with mmWave radar whose signals are reflected back from the human body and ambient objects. Our proposed mmMesh system in the paper aims to reconstruct the dynamic human mesh in real-time by taking the reflected mmWave signals as input. Figure 2 shows an overview of our proposed mmMesh system, which contains three major components: data collection, data preprocessing, and mesh construction.

Data Collection. This component aims to collect mmWave signals that can be used to reconstruct the subject's mesh. In this process, the commercial mmWave radar emits FMCW signals from its transmitting antennas and captures the reflected signals using its receiving antennas. Then the radar hardware can mix the received signals with the transmitted signal to obtain the IF (Intermediate Frequency) signals, which are the outputs of the mmWave radar. Note that a real-time data collection system is achieved by our UDP protocol based program to enable the dynamic human mesh construction. In addition to the collection of mmWave data, we also use the VICON motion capture system [1] to obtain high precision dynamic pose information of the subject, which is utilized to generate the ground truth human mesh that can be used to train the proposed deep learning model in our system.

Data Preprocessing. This component is designed to remove the noisy signals reflected from the static ambient objects, and then generate the 3D point cloud so that they can be fed to the proposed deep learning models. Specifically, we first calculate the heatmap using both range-FFT and doppler-FFT and cancel the signal energy from the static objects. Then we calculate the AoA (Angle of Arrival) of the signals in both azimuth plane and elevation plane. Based on the range information and the angle information, the locations of the 3D points can be easily estimated. The points' coordinates combined with other point features (e.g., point velocity) will be fed to our proposed deep learning model.

Mesh Construction. The goal of this component is to construct the dynamic human mesh from the point cloud generated by the data preprocessing component. In this component, we propose a novel deep learning model that can estimate 3D human mesh by simultaneously encoding the global and local structures of the 3D point cloud in spatial dimension as well as the structural transformation of the points in temporal dimension. The details of the proposed mmMesh model will be described in section 4. A real-time mesh rendering tool is also implemented in the developed system.

4 METHODOLOGY

In this paper, our goal is to construct dynamic 3D human mesh using sparse 3D point cloud data collected by the mmWave device. Our proposed mmMesh model should be able to tackle the following three challenges to achieve this goal.

The first challenge is the sparsity caused by the low resolution of the commercial mmWave radar device. In RF-Avatar [48], which is the only work using RF signals to estimate human mesh, there are 4 transmitting antennas and 16 receiving antennas assembled on a T-shape holder [2]. However, the commercial mmWave radar has only 3 transmitting antennas and 4 receiving antennas [13], which results in only a resolution of 15° for azimuth plane and a resolution of 60° for elevation plane. Due to the low resolution of the device, each frame of the collected data only contains hundreds of points, among which only dozens of points are correlated to the human body. As a consequence, we have to estimate the locations of thousands of vertices on the human mesh based on the information provided by only dozens of points, which is technically impossible. To address this challenge, we incorporate the Skinned Multi-Person Linear (SMPL) model [24] into our model as an additional constraint. SMPL is a generative 3D human body model which parameterizes the human mesh using a low-dimensional shape vector (to characterize the height, weight, and body proportions of human body), a pose vector (to characterize the deformation of the human mesh under motion), a global translation vector, and a binary gender parameter. SMPL can act as a strong constraint which allows us to use only 86 parameters to represent the whole 3D human mesh instead of directly estimating the location of each of the thousands of vertices. In addition, SMPL model can encode the anatomic prior knowledge of human body. For example, the length of one's arm span is roughly equal to one's height, and the human body and limbs tend to have a symmetric structure. These anatomic prior knowledge can help us produce a realistic human mesh.

Secondly, in real world, it is difficult to align 3D points precisely with their corresponding body parts due to: 1) the ambient points, which are generated from the RF signals reflected by the surroundings and can be mistakenly aligned with the body segments of the subject; 2) the error-prone natural of the RF signals; 3) the multi-path effect of the RF signals. Thus, it is challenging to correctly align the points in the point cloud and accurately construct the human mesh. To address this challenge, we first alleviate the influence of the ambient points. Specifically, in our model design, we filter the ambient points and consider only the points close to the subjects. Then, instead of directly learning the information from each single noisy point which may be affected by the error-prone nature of the RF signals, we propose to learn *the local structure of the point cloud*, i.e., the spatial relations between each 3D point and its neighboring points. For example, given a single point, it might be difficult to tell which part of human body the point belongs to. However, a 3D point on the subject's arm should be close to the other points on the arm or connected body segments, but far away from the points on the feet. Obviously, the spatial relation of the 3D points can provide us some knowledge about the shape and pose of the body segments,

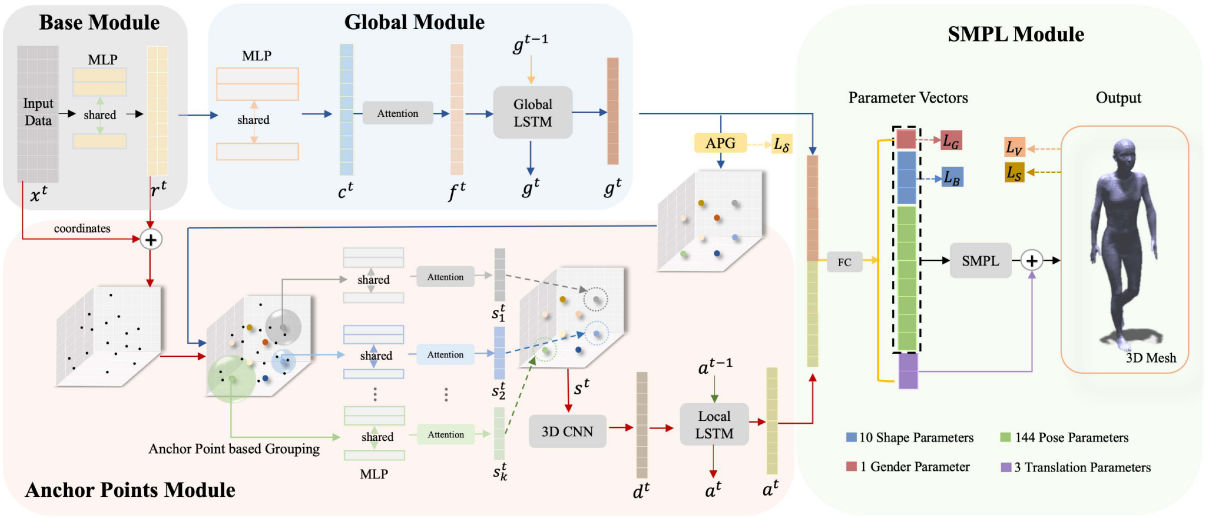

Figure 3: Model Overview

which is more robust to the error compared with the knowledge extracted from a single point. To capture the local structure of the point cloud, our proposed mmMesh model can *automatically group the neighboring points and associate point groups with corresponding body segments in a dynamic manner*. In addition, the qualities of different 3D points in the point cloud are usually quite different due to various reasons (e.g., the points generated from the RF signal reflected by the subject tend to have higher qualities than the ambient points; the points generated from the RF signals which are reflected multiple times in the space may have low qualities; the points that have high energy values in the corresponding Doppler-FFT heatmap tend to have high qualities). In our model design, we propose to use attention mechanism to discriminatively treat the points in the point cloud and automatically assign larger weights to the points that have higher qualities.

The third challenge is that some parts of the human body may not have correlated 3D points in a specific frame due to the specularity [47] of the RF signal. To address this challenge, we use the information (such as pose, shape and location of the subject) from previous frames to infer the missing body part information in the current frame. Specifically, we incorporate the Recurrent Neural Network (RNN) into our model to take advantage of the information from the previous frames.

Figure 3 gives an overview of our proposed mmMesh deep learning framework, which is mainly composed of four modules: a Base Module to extract the high-level representation of each point, a Global Module to aggregate the overall point cloud information, an Anchor Points Module to learn the local structure information of the point cloud, and a SMPL Module to map the generated representation vector to the final human mesh. The details of this framework will be described in the following subsections.

4.1 Base Module

The input of this module consists of the feature vectors of all the 3D points in the point cloud. In our work, one feature vector contains six features, which are x, y, z coordinates, the range value, the velocity value, and the energy value of the points in Doppler-FFT heatmap. These feature vectors are stacked into a 2-dimensional matrix as the input. In this module, the outputs are the high-level representations for all the points which are extracted by the share-weighted MLP (Multi-Layer Perceptron).

To be detailed, we use x^t to denote the input matrix of the t-th frame (obtained at time t), and use x_i^t to denote the feature vector of the i-th point in the input matrix x^t. The output of this module for i-th point in matrix x^t is denoted as $r_i^t = \text{MLP}(x_i^t; \theta_r)$, in which θ_r is the parameter set of the MLP layers.

4.2 Global Module

To locate the moving subject and estimate his/her shape and pose in a frame, we first need to extract the global information of the whole point cloud. The proposed Global Module can aggregate the information from all point representations derived by the Base Module, and combine them with the information from previous frames to form a global representation vector.

As shown in Figure 3, for each representation vector r_i^t derived by the Base Module, we first use MLP layers to map it into a higher-level vector representation $c_i^t = \text{MLP}(r_i^t; \theta_c)$. Then we aggregate all the point representations into a single vector. Note that the aggregation function should be permutation invariant to the order of the input, because the 3D point cloud is an unordered set. In other words, the aggregation function should output exactly the same output, no matter how the order of the input points are changed. In existing point cloud related work such as PoinNet/PointNet++ [27, 28], the max-pooling operation is usually used as the aggregation function to filter out redundant information and extract the most prominent features. However, in our scenario, since the point cloud is sparse and has little redundant information, using max-pooling operation may lead to the loss of some details in one frame. To address this problem, we adopt attention mechanism to aggregate the representations of all points in the current frame. Attention is the weighted sum of all point features without information loss and it allows us to dynamically learn relative contributions of each

point [15]. Suppose $L(x)$ denotes a linear mapping function that can map a vector into a scalar in attention operation. Then, we can get the aggregated global presentation f^t of all points in current frame as follows:

$$f^t = \sum_{i \in N^t} L(c_i^t; \theta_f) \cdot c_i^t, \quad (1)$$

where N^t is the number of points in current frame and θ_f is the parameter in the linear mapping function L. A key point here is that the attention function should be invariant to point permutation so that it can be applied to point clouds. The authors in [43] proved that a function $f(X)$ is invariant to the permutation of instances in X, *iff* it can be decomposed in the form of $\rho(\sum_{x \in X} \phi(x))$, for suitable transformations ϕ and ρ. Thus, our attention operation is invariant to point permutation, and it can be used to aggregate 3D point cloud features.

As aforementioned, some parts of the human body may not have correlated 3D points in a specific frame due to the specularity of the RF signal. We address this problem by leveraging the information of the previous frames to infer the missing parts. Specifically, we feed the representation vector f^t to the multi-layer LSTM and fuse it with previous global representations. Then we can get the final global representation of $g^t = \text{LSTM}(g^{t-1}, f^t; \theta_g)$, where g^{t-1} is the global representation of the previous frame and θ_g is a set of parameters to be learned in LSTM.

4.3 Anchor Point Module

We can get a rough estimation of the shape and pose of the subject by the Global Module. To make the estimated mesh more accurate, we need to learn the local structures of the point cloud to acquire fine-grained information. Traditionally, to learn the local structures of point cloud, the sampling method is first used to sample some points from the point cloud as grouping center. Then the points in the point cloud are grouped into several subsets. Finally, the representation of each subset is extracted and taken as the local structure representation [28].

However, the above method can only be applied to static objects and is not suitable to our scenario. The sampling strategy in the above method always samples points from the whole point cloud without distinguishing whether they are on the human body or not. There may be significant number of sampled points that are located far away from the human subject and thus contribute nothing but noise to the mesh construction. Moreover, The set of sample points are dynamically changing frame by frame and thus may lead to inconsistency across continuous frames.

To address this challenge, we propose to dynamically choose some "virtual locations" near the subject as *anchor points* and use them to group the 3D points. In our design, each anchor point can group a subset of points that are related to a part of human body. For example, the anchor points near the ground can group more points reflected from the calves of the human body, and the anchor points on the left of the subject may be more related to points on the left arm of the subject. Specifically, after deriving the global representation g^t from the Global Module, we use an Anchor Point Generator (APG) to generate the desired anchor points by taking g^t as input. In our design, the APG contains two phases: template generation and template displacement, as shown in Figure 4. In

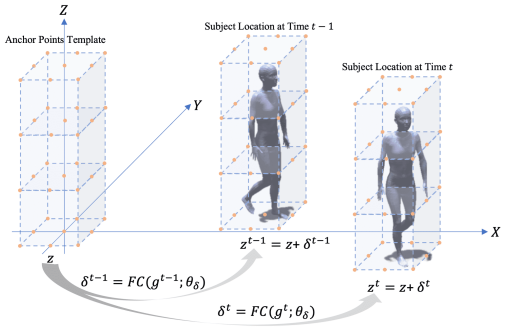

Figure 4: Anchor Point Generator (APG)

template generation phase, we first predefine an anchor point template at the origin as z, which is a 3D cubic lattice composed of N_z anchor points (red points in Figure 4) whose locations are fixed with respect to the anchor point template. In addition, we assume that the convex hull of the designed anchor point template is large enough to cover the subject. In template displacement phase, we first use a fully connected neural network (i.e., FC) to predict the displacement and then move the predefined anchor point template to the desired location. Here we use δ^t to denote the displacement of the template at time t, where $\delta^t = FC(g^t; \theta_\delta)$ is a coordinate vector with length 3. Then the anchor point template generated by APG at time t is located at $z^t = z + \delta^t$. Similar to the dynamic bounding box in object tracking task, the anchor point template can be dynamically generated at the locations of the moving subject frame by frame. This template can cover the 3D points of the human subject as many as possible, and meanwhile it is far from the points generated by the ambient objects.

As shown in Figure 3, based on the locations of the anchor points, we next group the 3D points from the Base Module into several subsets. One challenge here is that how some dynamic body segments (e.g., a swinging hand) can be captured by the anchor points with relatively fixed locations (with respect to the human body), since the points on these dynamic body segments can appear in the neighborhoods of different anchor points in different frames. In our model, though the anchor point template has fixed shape, the associations between the anchor points and the 3D points (on different body segments) are dynamic. And the dynamic associations are automatically learned during the point grouping and aggregation. Specifically, for each anchor point, we take the nearest N_s 3D points into a group. Suppose z_k^t denotes the location of the k-th anchor point at time t and the indexes of its nearest N_s points are represented as $NSP(z_k^t)$. We use $COOR(x_i^t)$ to represent the coordinate vector (i.e., a vector composed of x, y, z coordinate values) of point x_i^t in the Base Module. Then, for each point x_i^t where $i \in NSP(z_k^t)$, we can derive its high-level representation as $h_i^t = \text{MLP}([z_k^t; COOR(r_i^t) - z_k^t; r_i^t]; \theta_h)$, where θ_h denote the parameters of the MLP. Note that here we also encode the anchor point location (i.e., z_k^t) and the spatial relationship (i.e., $COOR(r_i^t) - z_k^t$) between the anchor point and its grouped points into the input of MLP. Similar to Eq. (1), the aggregation process based on the k-th

anchor point can be denoted as:

$$s_k^t = \sum_{i \in NSP(z_k^t)} L(h_i^t; \theta_s) \cdot h_i^t, \tag{2}$$

where θ_s denote the parameters of the linear mapping function L.

Next, we need to further aggregate the information in the anchor point representations. Since we carefully design the spatial relationship among the anchor points and arrange them as a 3D lattice in the cube, we can regard all the anchor point representation vectors as a 4D tensor s^t. Then we can aggregate the vectors of all the anchor points into one vector using 3D CNN as $d^t = 3DCNN(s^t; \theta_d)$, where θ_d denote the parametes in 3D CNN. Finally, similar to Section 4.2, information in the previous frames is fused with d^t using multi-layer LSTM as $a^t = \text{LSTM}(a^{t-1}, d^t; \theta_a)$, where θ_a denote the parameters of LSTM.

4.4 SMPL Module

In this module, we first concatenate the global representation vector from the Global Module and the local representation vector from the Anchor Point Module, and then map them into pose, shape, translation and gender representation vectors. Finally we feed the vectors into SMPL model to output the skeleton and meshe of the subject.

Specifically, a multi-layer fully connected neural network is used to get the representations as following:

$$[P^t; B^t; T^t; G^t] = FC([g^t; a^t]; \theta_p),$$

where P^t is the pose vector, B^t the shape vector, T^t is the translation vector, and G^t is the gender vector. Note that in original SMPL paper, the length of the pose vector is 72 and it is composed of 24 rotation vectors. However, according to [50], 3D rotation vector is not a continuous rotation representation to neural network. Thus, following [50], we use 6D representation to represent the rotation. And the length of the pose vector P^t in our model is $144 = 24 \times 6$. Then the vertex vector V^t and skeleton vector S^t can be obtained by feeding the P^t, B^t and G^t into the SMPL model as following:

$$[V^t; S^t] = \text{SMPL}(P^t, B^t; G^t) + T^t.$$

Note that the mesh models for male and female are different. Our SMPL Module can automatically select the corresponding mesh model based on gender vector G^t. Since SMPL model only takes the 3D rotation vectors as input, we implement a function inside the SMPL model to transform the 6D rotation representations to the 3D rotation vectors. In addition, the parameters of SMPL model are trained in [24] and keep freezing in our model.

4.5 Model Loss

The model loss is the summation of 5 components as following:

$$\begin{aligned} Loss = \sum_{K \in \{V,S,B,\delta\}} \alpha_K * \sum_t^T \|K^t - \mathcal{GT}(K^t)\|_{L_1} \\ + \alpha_G * \sum_t^T H(G^t, \mathcal{GT}(G^t)). \end{aligned} \tag{3}$$

Here we use V, S, B, G to denote the vertex matrix, skeleton matrix, shape matrix, and gender matrix obtained in the SMPL module from the first frame to the T-th frame. δ is the displacement matrix obtained using APG in the Anchor Points Module. We use $\mathcal{GT}(K)$ to denote the corresponding ground truth of the generated matrix K and α_K denote the hyper-parameters. H is the hinge loss. Normally, cross entropy will be used to classify the gender of the subject. However, the cross entropy loss can be very large, which may affect other losses. To address this problem, we use hinge loss on the gender vectors. Note that even though the vertex loss is the joint result of pose, shape, displacement, and gender of the subject, we still add the skeleton loss, shape loss, displacement loss, and the gender loss to guide the fast convergence of the designed deep model and to avoid the model falling into the local minimum.

5 EXPERIMENTS

5.1 Testbeds

5.1.1 VICON System. In this paper, we use the VICON motion capture system [1] to generate ground truth 3D human pose for model training. The VICON system is shown in Figure 5(c), and it consists of 21 VICON Vantage cameras which emit and receive infrared light. During the pose data collection, 27 high precision pearl markers are placed on each subject to represent the joint points of the subject. Figure 5(a) shows the positions of these markers on the subject. Since these markers are covered with highly reflective materials, the infrared light reflected from the marker surface can be easily captured by the the VICON Vantage camera. The errors cause by the location of each marker is less than $2mm$ [25]. The sampling rate of the system is 10 frames per second.

5.1.2 mmWave Testbed. The millimeter-wave radar we used in this paper is TI AWR1843BOOST, which is a commercial and portable ($8.3cm \times 6.4cm$, $30g$) mmWave device produced by TI [13]. We also utilize TI DCA1000EVM to enable real-time data capture and streaming from mmWave radar, as shown in Figure 5(b). The mmWave device contains 3 transmitting antennas and 4 receiving antennas. The 3 transmitting antennas emit FMCW wave chirps in turns. The emitted RF singal will be reflected by the human body and the surroundings, then received by the 4 receiving antennas. For each FMCW chirp, the frequency of RF will increase from 77 GHz to 80.9 GHz. The mmWave device is set to send 10 frames per second. Here each frame is composed of 128 chirps, and each chirp is composed of 256 sampling points. Based on our device setting, the maximum sensing range of the mmWave device is about 11 m, the range resolution is about 4.3 cm, the maximum sensing velocity is about 4.5 m/s, and the velocity resolution is about 7.1 cm/s. To enable a real-time system, a UDP-based program is developed to collect packets from the device and parse the packets into the mmWave data frame. In the experiment, we place the mmWave testbed on a table (the height is about 92 cm), and the distance between the mmWave testbed and the activity area is about 1.5 m.

5.2 Data Collection and Preprocessing

5.2.1 Data Collection. In the experiment, 20 volunteers (including 13 males and 7 females) are asked to perform 8 daily activities within the activity area. The 8 activities include: (1) torso rotations; (2) clockwise walking; (3) counter-clockwise walking; (4) arm swing (the subject can randomly swing his/her arms horizontally or upward or downward); (5) walking back and forth; (6) walking back

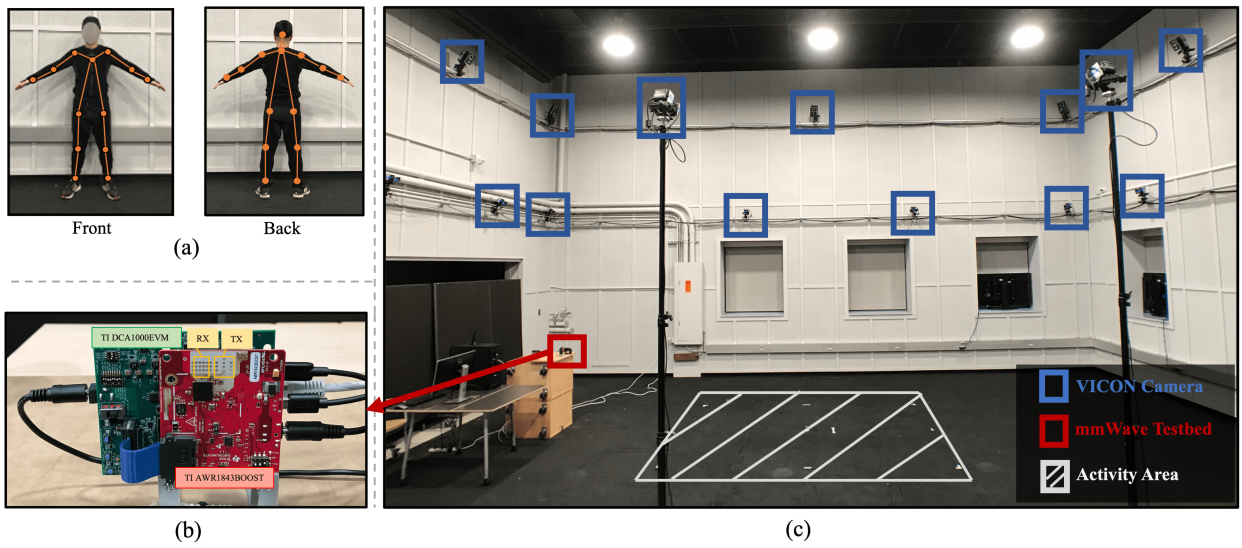

Figure 5: Testbeds and the basic scenario of mesh construction

and forth with arm swing; (7) walking in the place; (8) lunges (the subject keeps performing lunge pose alternatively use his/her left and right leg). For each activity, the subject keeps performing it for 5 minutes (i.e., 3000 frames per activity per subject).

5.2.2 Ground Truth Mesh Construction. In our experiment, we use SMPL model to generate the ground truth human mesh to train our proposed deep learning model. Specifically, we take the pose information, shape information, translation information, and gender information of the subjects as the input of SMPL model.

The pose information and translation information can be obtained from the VICON system as described in Section 5.1.1. Note that the pose representations obtained from the VICON system are the absolute positions of the joints. As shown in Figure5(a), the location of each joint is obtained by averaging the locations of the two markers that are nearest to the joint (the two markers are in the front and the back of the human body, respectively). Since the pose representations are the rotation vectors of the joints, we then calculate the rotations on the joints using the absolute positions obtained from the VICON system. It is notable that the joints from VICON system and the input pose vectors of SMPL model are not one-to-one mapping, the SMPL model has more pose vectors than the joints from VICON system. Since the missing joints have little effect on the designed daily activities, we simply set those rotation vectors with constant values.

For the shape information, we use the approach in [5] to obtain the ground truth shape vector for each subject in a canonical pose. To best match the human mesh model with the ground truth height of the subject, we also manually adjust the shape vector values.

5.2.3 Point Cloud Generation. After obtaining the frames of the mmWave data, we first calculate the Range-FFT and Doppler-FFT. Then a static clutter removal algorithm is used to remove the static background noise. The algorithm subtracts the average value of the Doppler-FFT heatmaps from all receiving antennas, which is helpful to reduce the energy reflected from static ambient objects.

Traditionally, CA-CFAR [8] algorithm is usually applied on the Doppler-FFT heatmap to select prominent pixels as potential 3D points using fixed threshold. However, in our work, we directly use the Doppler-FFT heatmap pixels with the highest values as the potential 3D points. The main reason is that the heatmap energy differs from frame to frame. Using the fixed threshold will result in the number of selected points varying largely in different frames. For example, if we use the fixed threshold, one frame may have a hundred of selected points while another frame may have only several or even zero selected points. This effect will be severe when the testing environment changes. Especially when we conduct experiment in occlude scenarios as described in Section 5.5.4, the energy distribution changes largely when the signal from the device is occluded by objects. Though we can manually set the threshold for each environment, we choose to select 128 heatmap pixels with the highest values for each frame to generate consistent numbers of point clouds among different frames and various environments. The noisy points generated by the multi-path effect is also alleviated in this step, since the signal that has been reflected several times tends to have lower energy value than the directly reflected signal. Then, we calculate the 3D coordinates of the points based on the the selected pixels. Finally, we take the x-y-z coordinates, the range value, the velocity value, and the Doppler-FFT value of the point as the input feature vector of each 3D point to feed to the deep learning model.

5.3 Model Setting and Model Training

In this section, we describe the setting of the deep learning model. In Base Module, we use 3 layers of shared MLPs and the sizes of the layers are 8, 16, and 24, respectively. In Global Module, we also use 3 layers of shared MLPs and the sizes of the layers are 32, 48, and 64, respectively. The LSTM in Global Module has 3 layers and the size of each layer is 64. In Anchor Points Module, we use 81 anchor points which construct a $3 \times 3 \times 9$ 3D cubic lattice. The distance between a pair of neighboring anchor points is set to 0.3 m. The

number of grouped points around each anchor point is set to 8. The size of the shared MLPs in Anchor Points Module is the same as that in Global Module. We use 3 layers of 3D CNN to aggregate the featueres of anchor points into a vector with size 64. Similarly, there are 3 layers of LSTM in Anchor Point Module whose sizes are all set to 64. The FC has 2 layers in SMPL Module, which maps the concatenated vectors into a parameter vector with size 158.

During model training, we use the first 2400 frames (i.e., 80% of the data) of all subjects' activities for training, and the remaining 600 frames (i.e., 20% of the data) for testing. The learning rate is set to 0.001. The batch size is 32. The sequence length for training is 64. The number of training batches is 500 K. The weights assigned to different losses in Eq. (3) are set to $\alpha_V = 0.001$, $\alpha_B = 0.1$, and $\alpha_S = \alpha_\delta = \alpha_G = 1.0$. Note that the ground truth gender is used to select the mesh model in the SMPL Module during the training. However, during testing, we only use the predicted gender to select the mesh model. We use PyTorch to implement our deep learning model, and TITAN V is used to train the model.

5.4 Baselines and Metrics

5.4.1 Baselines. Since there is no existing model to reconstruct dynamic human mesh from point clouds, we design the baselines by removing or replacing the modules in the architecture of the proposed model as following:

B+G+S (Baseline A). In this baseline, the Anchor Point Module is removed. We inherit the Base Module, the Global Module, and the SMPL Module from the proposed mmMesh model without any change.

B+G-Max+S (Baseline B). This model shares the same structure with *Baseline A* except that the attention-based grouping is replaced with max-pooling operation in Global Module.

B+G+FPS-ATTN+S (Baseline C). In this baseline, besides inheriting the Base Module, Global Module, and SMPL Module from the proposed mmMesh model as in *Baseline A*, we use FPS-based sampling [28] and aggregate the features of grouped points using attention mechanism.

B+G+FPS-Max+S (Baseline D). This model is very similar to *Baseline C* except that we replace attention-based aggregation method with the max-pooling operation. For this baseline, the model design to learn the local structure is the same as that of PointNet++ [28] which utilizes the FPS-based sampling and max-pooling operation based aggregation method.

5.4.2 Metrics. We use the the following metrics to evaluate the performance of our proposed framework:

Average Vertex Error (V) [6, 48]. We compute the average vertex error by averaging the Euclidean distance between the vertices located on the predicted human mesh and the corresponding vertices on the ground truth mesh for all the subjects and activities. This metric can evaluate the overall performances of the location error, pose error, shape error, and gender error.

Average Joint Localization Error (S) [16, 48]. This metric is defined as the average Euclidean distance between the joint locations of the predicted human mesh and the ground truths for all the subjects and activities.

Average Joint Rotation Error (Q). Besides the joint position, joint rotation is also critical when generating the pose. This metric is reported as an additional metric to evaluate the accuracy of the constructed pose. It is defined as the average differences between predicted joint rotations and the ground truth rotations. As described in Section 5.2.3, some joint rotations in SMPL model are set to constant values. There is no need to take these joints into consideration. Thus, when calculating the average joint rotation error, we only consider the rotations of shoulder joints, elbow joints, hip joints, and the knee joints from both sides of the subject.

Mesh Localization Error (T). We also use mesh localization error to assess the precision of subject localization. This metric is defined as the average Euclidean distance between the root joint location of the predicted human mesh skeleton and the ground truths for all the subjects and activities.

Gender Prediction Accuracy (G). We also calculate the accuracy of the predicted gender to evaluate if the proposed model can distinguish gender of the subject.

5.5 Experiment Results

5.5.1 Qualitative Results for Basic Scenario. We first qualitatively evaluate the proposed framework in the basic scenario that is shown in Figure 5(c). The setting of the training phase for the basic scenario is described in Section 5.3. The qualitative results are shown in Figure 6, in which rows (a)-(c) show 3 male subjects conducting activities 8, 5, and 4, respectively. Rows (d)-(f) show 3 female subjects conducting activities 1, 6, and 7, respectively. As we can see, the six subjects in this figure have different shapes. The first picture in each row shows the video frame when the subject conducting the activity. The second picture and the third picture show the corresponding ground truth human mesh generated by the VICON system and the predicted human mesh based on our proposed mmMesh model, respectively. The results show that our generated meshes look realistic. From this figure, we can see the shapes of the generated human meshes are very similar to the corresponding subjects in the video frames. In addition, our model can predict the correct gender of each subject, which demonstrates that our model is able to correctly sense the gender information of the subjects and generate the human meshes with reasonable shapes, even if the subjects in our experiment have different heights and shapes. The results in this figure also show that our proposed mmMesh model can accurately estimate the human poses. This demonstrates that our model is able to capture the subtle body structure information from the local structure of the point cloud.

The rows (g)-(l) in Figure 6 show the consecutive frames that one subject is conducting clockwise walking (activity 2) within the activity area. The first, second, and third columns show the video frames, ground truth meshes, and the human meshes generated by our model, respectively. To better show the activity process, we pick every other frame in the video (i.e., the time gap between each pair of consecutive frames is $0.2s$). We can see that the constructed mesh in consecutive frames looks not only very similar to the ground truth meshes, but also very smooth. This is achieved by taking advantage of LSTM layers in our model, which encodes the temporal information in the network. This result proves that our model can generate smooth dynamic human meshes.

5.5.2 Quantitative Results for Basic Scenario. In this section, we quantitatively evaluate the performance of our proposed model

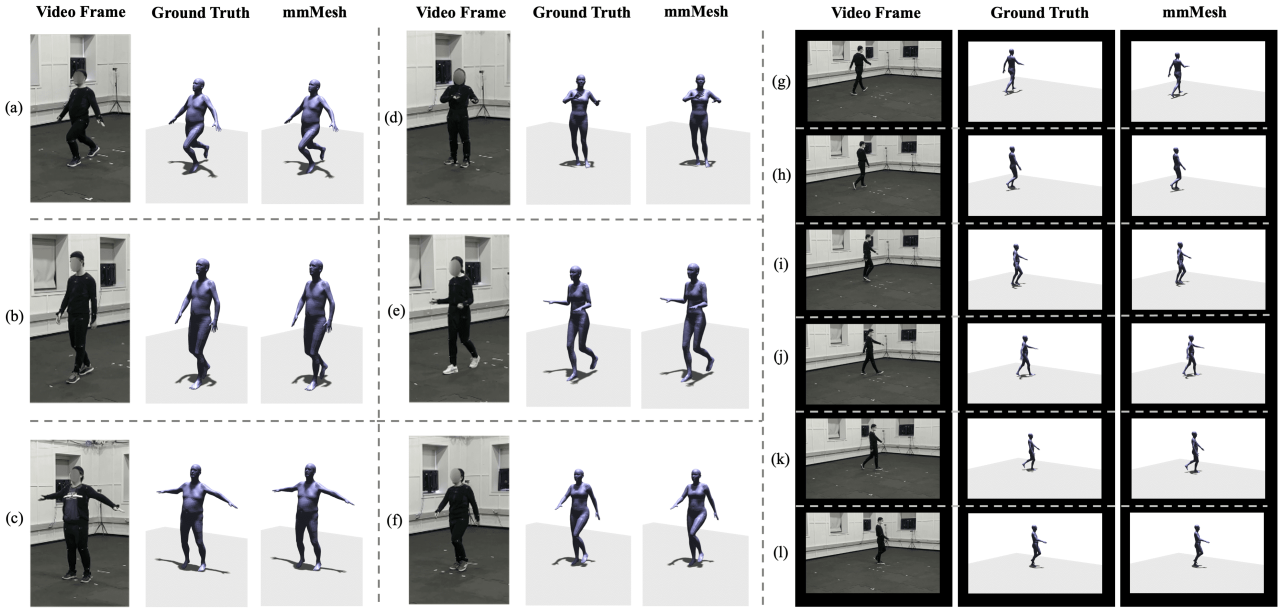

Figure 6: The examples of the constructed human mesh in the basic scenario.

Table 1: Results for Basic Scenario.

Model	V(cm)	S(cm)	Q(°)	T(cm)	G(%)
Baseline A	3.75	3.43	4.57	2.42	99.5
Baseline B	3.88	3.58	4.74	2.54	99.0
Baseline C	3.43	3.10	4.26	2.16	99.2
Baseline D	3.63	3.29	4.42	2.32	97.2
mmMesh	**2.47**	**2.18**	**3.80**	**1.27**	**99.8**

based on the metrics described in section 5.4.2. The results are shown in Table 1. As can be seen, for all five metrics utilized in this paper, our proposed mmMesh model achieves the best results. This demonstrates that our model is able to generate more accurate poses, shapes, genders and translations of the subjects than all the baselines.

To study the difference between the attention-based aggregation method and the max-pooling-based aggregation method, we compare the performance of baselines A and B as well as that of baseline C and D. Baseline A and baseline B share the same structure except that baseline A uses attention mechanism while baseline B uses max-pooling operation. Similarly, Baseline C and baseline D share the same structure except that baseline C uses attention mechanism while baseline D uses max-pooling operation. As we can see in the Table 1, although the gender accuracy of baselines A and C is slightly worse than that of baselines B and D, respectively, baselines B and D perform better than baselines A and C on other metrics. This means the overall performance of the proposed model is improved by replacing max-pooling operation with attention mechanism. This is because the point clouds in our scenario are very sparse. Using max-pooling operation may cause the model insensitive to subtle structures of the point cloud and impair the model performances. As a substitution, the attention mechanism is able to distinctively sum up the point representations and aggregate them with little information loss.

Next, we study the importance of learning local structures of point cloud to the construction of human mesh. In baselines A and B, there are no design to learn local structures of point cloud. But in baselines C and D, we use FPS-based sampling to learn the local structures. In addition, our proposed mmMesh model use anchor point based method to learn local structures. From Table 1 we can see that the models (i.e., baselines C and D and our proposed model) with the design to learn local structures perform better than those (i.e., baselines A and B) without learning local structures, even baseline A uses attention mechanism and baseline C use max-pooling operation. This is mainly because we can capture more detailed information about the human body structure by learning local structures of point cloud.

The results in Table 1 also show that our proposed model outperforms baseline C and the performance on metrics V, S and T are all improved by about 1 cm. This is because our model uses anchor point based sampling method while baseline C uses FPS-based sampling. The anchor point sampling method can dynamically sample the points near the subject and avoid including the noisy points from the ambience.

We also evaluate the performance of our model with different training rates of the data. Specifically, we vary the training rate from 50% to 80%, and the results are shown in Table 2.It can be seen that the performance of our model only has a small drop when the training rate is reduced from 80% to 50%, which demonstrates the robustness of our model.

5.5.3 Performance Evaluation for Occluded Scenario. To investigate the effect of occlusion on the performance of our proposed mmMesh

Table 2: Results for Different Training Rates

Training Rate	V(cm)	S(cm)	Q(°)	T(cm)	G(%)
50%	2.89	2.54	4.40	1.47	99.1
60%	2.76	2.42	4.18	1.41	99.5
70%	2.65	2.34	4.00	1.39	99.5
80%	2.47	2.18	3.80	1.27	99.8

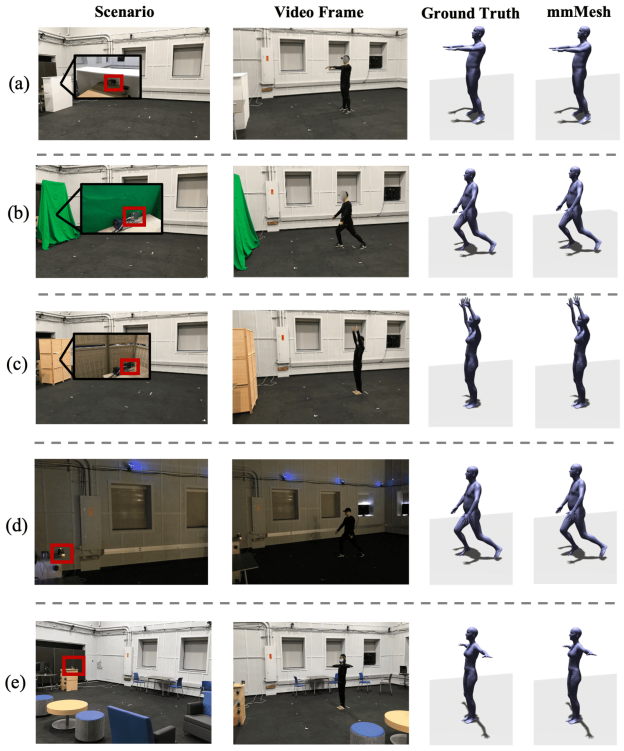

Figure 7: The examples of the constructed human mesh in occluded scenarios and cross-environment scenarios. (The mmWave radar is marked inside the red box)

framework, we place barriers of different material between the subject and the mmWave radar. As shown in Figure 7, the first column of rows (a), (b), and (c) show three occluded scenarios, where we use a foam box, a cloth screen, and a bamboo panel as the barriers, respectively. For each scenario, the VICON system is adopted to collect the ground-truth poses. In this experiment, we ask 5 subjects to perform the 8 activities for 2 minutes. In Figure 7, the second, third, and fourth columns show the video frame, the corresponding ground truth human mesh, and the human mesh constructed by our model, respectively. Note that in these three occluded scenarios, we directly use the trained mmMesh model from the basic scenario during the inference. As we can see, our model can still generate high quality human mesh with accurate pose and shape, even the transmitted signal is completely occluded by different barriers.

Table 3: Results for different occluded scenario.

Occluded Scenario	V(cm)	S(cm)	Q(°)	T(cm)	G(%)
Foam Box	5.93	5.54	8.35	3.88	96.8
Cloth Screen	6.33	5.87	8.88	3.88	96.8
Bamboo Panel	6.45	6.06	8.67	4.57	87.4

Table 4: Results for the room with different settings.

Room Settings	V(cm)	S(cm)	Q(°)	T(cm)	G(%)
Dark scenario	5.47	5.14	7.91	3.13	97.3
Furnished	5.95	5.53	8.27	3.93	94.4

We also quantitatively study the performances of our model in the occluded scenarios. Table 3 reports the results using the five metrics that are described in Section 5.4.2. By comparing the results for basic scenario in Table 1, we can see that the occlusion degrades the performance of our proposed model. However, the human mesh with high quality can still be constructed. There are mainly two reasons for the increase of the errors. One reason is that the signal phase is changed when the signal penetrates the barriers, which can affect the location accuracy of the points. Since the material of bamboo has the most compact structure, we can see that the bamboo panel has the largest effect on the model performance and the foam has the smallest effect. The other reason is that the systematic error may be introduced during the re-calibration of the VICON system and re-adjusting of markers. Note that the data in the basic scenario and that in the occluded scenario are collected on different dates. The VICON system need to be re-calibrated each time we use it. Since the coordinated system are labelled using the calibration wand manually, some errors may be introduced in this step. What's more, during the data collection, we need to re-adjust the locations of the markers attached on the suit manually, this step can also introduce some errors.

5.5.4 Performance Evaluation for Cross-environment Mesh Construction. Another challenge when using our mesh reconstruction system in real world is that how to make it adapt to different environments. As aforementioned, mmWave signals can be reflected by the objects in the ambient environment. Different objects in different environments may cause different ways of transmitting of the signals.

In order to investigate the effect of environment changing on the performance of our system, we first conduct experiment in the room with the VICON system as illustrated in rows (d) and (e) of Figure 7. Row (d) shows a dark scenario, in which the vision-based methods usually have poor performance. Row (e) shows a furnished scenario, in which the furniture (e.g., tables and chairs) is randomly placed around the activity area. The quantitative results for the two scenarios are reported in Table 4. In this experiment, we still directly use the model trained in the basic scenario without additional training. The ground truth poses are collected using the VICON system.

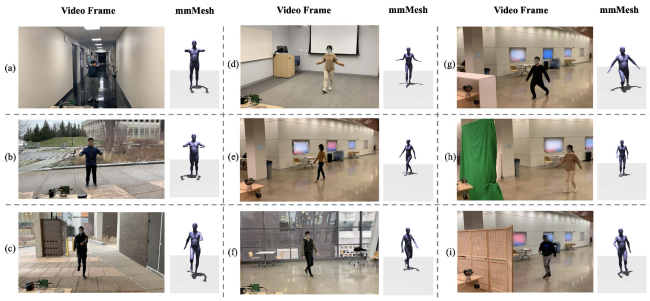

Figure 8: The examples of the constructed human mesh in different environments

From the results we can see that even the environment is changed, our model can still construct good human mesh. The results in the dark scenario demonstrate the advantage of RF-based sensing solutions over the vision-based sensing solutions when the light condition is bad. Note that the performance here is not as good as that in basic scenario. The reason is similar to that for the occluded scenario. The re-calibration of the VICON system and re-adjusting of markers can introduce systematic errors. For the furnished scenario, our model can also output high quality human meshes, which means the anchor point module is able to filter out the ambient noisy points by dynamically shift anchor point template with the locations of the subjects.

In addition, to demonstrate the feasibility of our model in real practice, we further evaluate it in some more challenging scenarios. Specifically, we evaluate our model in completely new scenarios as illustrated in Figure 8. This figure shows the subjects conducting different activities in different environments and the corresponding human mesh generated based on our model. These environments include a corridor (a), an outside plaza (b), a passage beside a building (c), a meeting room (d), a hallway inside a building (e), and a student lounge. The rows (g) - (i) in Figure 8 show more challenging scenarios, where the barriers (foam box, cloth screen and bamboo panel) mentioned in Section 5.5.3 are placed between the subject and the mmWave radar in the hallway scenario (e). The corresponding model outputs are placed next to the video frames, which are obtained by directly using the trained model in the basic scenario. In this figure, the angles of the camera may vary when placed in different environments, and the output meshes are rendered from the perspective of the mmWave radar. Note that due to the absence of the VICON system in these scenarios, we cannot generate the ground truth meshes. However, the results in Figure 8 show that the poses and the shapes of the generated meshes are very similar to that of the subjects in the video frames, which demonstrates the effectiveness of our model in different environments.

It is worth mentioning that during our experiments, we found that our model may estimate inaccurate human shape at the very beginning of inference. As shown in Figure 9, Figure 9(a) shows the first frame of a video, Figure 9(b) shows the corresponding estimation result, and Figure 9(c) shows the ground truth shape of the subject in Figure 9(a). We can see that the estimated human shape in Figure 9(b) is somehow different from the ground truth shape of the subject in Figure 9(c). The reason for the inaccurate inference is

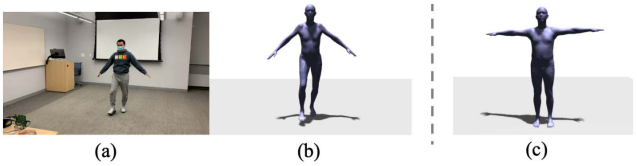

Figure 9: An inaccurate shape estimation case: (a) the first frame in a video; (b) the estimated mesh; (c) the ground-truth shape.

that the point cloud generated by the commercial mmWave radar is too sparse to enable an accurate shape estimation from one single frame. Our model tackles this problem by incorporating the information from the previous frames. However, there is no referable historical information at the very beginning of the inference. Thus, there might be inaccurate shape estimation at the very beginning of the inference.

5.6 Real-time System Implementation

In this paper, we aim to achieve real-time human mesh reconstruction. When collecting the data, the RF signals are transmitted and received by TI AWR1843BOOST mmWave device, and the attached TI DCA1000EVM enables the real-time IF signal translation using UDP protocol. In our design, we write a program to unpack the UDP packets and assemble them into the data frames. Note that a data frame will not be passed to the next step until it is completely assembled. Theoretically, the total mount of floating-point arithmetic of our whole model is $11.4M$ per frame, which is quite small. In practice, the time delay in the data collection process is about 110 ms. The assembled data frames are then fed into the data preprocessing component (as shown in Figure 2) to generate the point clouds. For each frame, the preprocessing time is about 28 ms using Intel i7-8700K CPU. After generating the point clouds, we next feed them into NVIDIA GTX1080Ti GPU for deep model inference. The inference time of the mmMesh model is about 16 ms. Finally, the 3D human meshes are rendered using the same GPU and the time delay is about 3 ms. Note that to make the programs more flexible, each program has its own clock to fetch the data from the previous step, which introduces a time delay about 100 ms. Thus, the total time delay in our system is about 257 ms. Figure 10 shows the consecutive frames of human mesh generated by our system. We can see that the rendered meshes in (d), (e), and (f) are roughly corresponding to the subject poses in (a), (b), and (c) respectively, which demonstrates that our system can achieve almost a real-time mesh reconstruction with a delay within about 3 frames (0.3s).

6 RELATED WORK

mmWave-based Sensing: mmWave has been increasingly explored to enable various sensing tasks, especially the human sensing tasks, such as human monitoring and tracking [3, 38, 44], pose estimation [21, 30, 31], behavior detection and recognition [19, 23, 32, 46], human acoustic sensing [22, 39], and human detection and identification [9, 41]. Being different from all the above tasks, our work takes a step forward and aims to construct dynamic human mesh using the mmWave signals in real-time. In addition to human

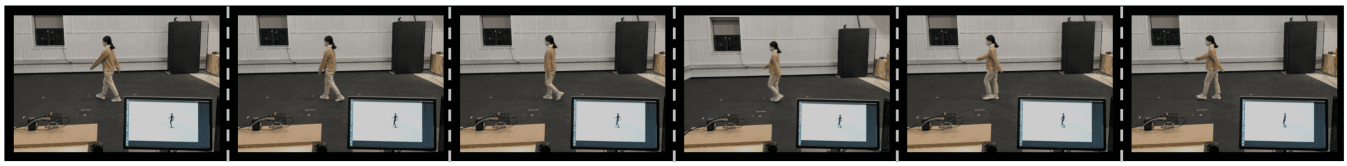

Figure 10: The examples of the constructed human mesh in consecutive video frames

sensing, mmWave is also utilized on industrial vibration measurement tasks [14] and car imaging tasks [10]. However, the proposed methods are only suitable to sense stationary objects. Since the subject in human sensing is always dynamic, those methods can not be directly applied to human sensing tasks.

Human Pose Estimation from Wireless Signals: In recent years, many wireless sensing systems have been developed to estimate human pose [2, 16, 34, 35, 47, 49]. Among them, [34, 35, 47] focus on 2D pose estimation. RF-Pose [49] can estimate 3D multi-person pose. However, the method requires specially designed testbed with an carefully assembled and synchronized antenna array. Most recently, Jiang et al.[16] propose WiPose to construct 3D human skeletons from WiFi signals. However, WiPose requires that the locations of the subjects should be fixed. In the above human pose estimation works, none of them can achieve real time estimations. In addition, all the wireless devices used in these works are discommodious to move. In our paper, the proposed mmMesh system can not only generate the human mesh as an enrichment of the human pose, but also be implemented in a real-time manner. It is worth mentioning that in our system design, we choose the commercial and portable mmWave device instead of the common WiFi device. The main reason is that the common WiFi devices do not use FMCW signal which enables accurate measurements (e.g., ToF) of RF signals, and thus cannot achieve as good performance as mmWave radars.

3D Human Mesh Construction: With the proliferation of deep learning, recent works explore various deep learning models to directly reconstruct 3D human mesh from images [6, 17, 20, 26, 33, 42, 51], videos [4, 11, 18, 33, 40, 45], point cloud [12, 15, 36, 37], and wireless signals [48]. Despite the great success achieved by image/video based approaches, the performance of these methods can be severely impaired by bad illumination, occlusion and blurry. Most importantly, privacy issues occur when cameras are deployed to monitor the human subjects. In contrast, our mmWave based approach can not only avoid the privacy issue but also be immune to the poor lighting and occlusion conditions. Since our proposed method reconstruct 3D human mesh from point cloud collected with mmWave radar, here we mainly introduce the previous works that utilize point cloud or wireless signals for human mesh reconstruction.

Point Cloud based: Recently, with the rapid development of 3D point cloud acquisition technologies, 3D human mesh construction from point cloud have been attracting more and more attention [7, 15, 36, 37]. As pioneering models for point cloud feature learning, PointNet [27] and PointNet++ [28] are widely used as the basic block to develop other methods for 3D human mesh reconstruction. And the parametric human body model (e.g., SMPL [24]) is also used in [15, 36, 37] for point cloud based human mesh reconstruction. However, these solutions only focus on the clean 3D point cloud without noise points from the ambience. And the point cloud contains thousands of point from the whole human body. Thus, these solutions cannot be directly applied in our scenario.

Wireless Signal based: As far as we know, there is only one work that have explored 3D human mesh reconstruction using wireless signal. RF-Avatar [48] first obtains a 4D RF tensor from its FMCW radios, and this 4D RF tensor is composed of many 3D energy distribution tensors arranged along the time dimension. Then, the proposal network, self-attention mechanism, and adversarial training are applied on the 4D RF tensor to output the 3D human mesh sequence. Our designed system is different from RF-Avatar. First, RF-Avatar is based on a specialized testbed which is a carefully assembled and synchronized USRP-based bulky antenna array [2], and this limits its real-world deployments. In contrast, our proposed system can accurately estimate the human mesh by using only a commercial mmWave device that can be directly purchased online at a low cost. Second, the model design in [48] can only be applied to the devices based on which the energy map of the 3D space can be obtained. However, our proposed model can be deployed on any devices that can generate 3D point clouds (e.g., Kinect, LiDAR, and depth camera), which enables a wide spectrum of real-world applications. Third, our system is able to directly estimate the dynamic human mesh in a real-time manner while [48] can only perform the evaluation offline due to its model design.

7 CONCLUSIONS

In this paper, we study how to use mmWave signals to construct dynamic human mesh in real-time. Specifically, we propose a deep learning framework, named mmMesh, which can construct human mesh using the point cloud generated from mmWave signals. This framework encodes a 3D human body model to tackle the sparsity of the point cloud. It also incorporates an anchor point module to handle the misalignment of the point cloud with the human body segments and leverages the information from the previous frames to address missing body parts problem. In addition, we implement a prototype of our mmMesh system using COTS millimeter wave devices. The evaluation results show that our mmMesh system can accurately localize the vertices on the human mesh.

ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation under Grants IIS-1938167, OAC-1934600 and CNS-1652503. And we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

REFERENCES

[1] [n.d.]. VICON Motion Systems. https://www.vicon.com.

[2] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the human figure through a wall. ACM Transactions on Graphics (TOG) 34, 6 (2015), 1–13.

[3] Mostafa Alizadeh, George Shaker, João Carlos Martins De Almeida, Plinio Pelegrini Morita, and Safeddin Safavi-Naeini. 2019. Remote monitoring of human vital signs using mm-Wave FMCW radar. IEEE Access 7 (2019), 54958–54968.

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Detailed human avatars from monocular video. In 2018 International Conference on 3D Vision (3DV). IEEE, 98–109.

[5] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video based reconstruction of 3d people models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 8387–8397.

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In European Conference on Computer Vision. Springer, 561–578.

[7] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. 2020. Implicit functions in feature space for 3d shape reconstruction and completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6970–6981.

[8] HM Finn. 1968. Adaptive detection mode with threshold control as a function of spatially sampled clutter-level estimates. RCA Rev. 29 (1968), 414–465.

[9] Tianbo Gu, Zheng Fang, Zhicheng Yang, Pengfei Hu, and Prasant Mohapatra. 2019. mmSense: Multi-Person Detection and Identification via mmWave Sensing. In Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems. 45–50.

[10] Junfeng Guan, Sohrab Madani, Suraj Jog, Saurabh Gupta, and Haitham Hassanieh. 2020. Through Fog High-Resolution Imaging Using Millimeter Wave Radar. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11464–11473.

[11] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2019. Livecap: Real-time human performance capture from monocular video. ACM Transactions on Graphics (TOG) 38, 2 (2019), 1–17.

[12] Taisuke Hashimoto and Masaki Saito. 2019. Normal Estimation for Accurate 3D Mesh Reconstruction with Point Cloud Model Incorporating Spatial Structure.. In CVPR Workshops. 54–63.

[13] Texas Instruments. [n.d.]. . http://www.ti.com.

[14] Chengkun Jiang, Junchen Guo, Yuan He, Meng Jin, Shuai Li, and Yunhao Liu. 2020. mmVib: micrometer-level vibration measurement with mmwave radar. In Proceedings of the 26th Annual International Conference on Mobile Computing and Networking. 1–13.

[15] Haiyong Jiang, Jianfei Cai, and Jianmin Zheng. 2019. Skeleton-Aware 3D Human Shape Reconstruction From Point Clouds. In Proceedings of the IEEE International Conference on Computer Vision. 5431–5441.

[16] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D human pose construction using wifi. In Proceedings of the 26th Annual International Conference on Mobile Computing and Networking. 1–14.

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7122–7131.

[18] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. 2019. Learning 3d human dynamics from video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5614–5623.

[19] Soo Min Kwon, Song Yang, Jian Liu, Xin Yang, Wesam Saleh, Shreya Patel, Christine Mathews, and Yingying Chen. 2019. Hands-Free Human Activity Recognition Using Millimeter-Wave Sensors. In 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN). IEEE, 1–2.

[20] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. 2017. Unite the people: Closing the loop between 3d and 2d human representations. In Proceedings of the IEEE conference on computer vision and pattern recognition. 6050–6059.

[21] Guangzheng Li, Ze Zhang, Hanmei Yang, Jin Pan, Dayin Chen, and Jin Zhang. 2020. Capturing Human Pose Using mmWave Radar. In 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). IEEE, 1–6.

[22] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. 2020. VocalPrint: exploring a resilient and secure voice authentication via mmWave biometric interrogation. In Proceedings of the 18th Conference on Embedded Networked Sensor Systems. 312–325.

[23] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kunpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. 2020. Real-time Arm Gesture Recognition in Smart Home Scenarios via Millimeter Wave Sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4, 4 (2020), 1–28.

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG) 34, 6 (2015), 1–16.

[25] Pierre Merriaux, Yohan Dupuis, Rémi Boutteau, Pascal Vasseur, and Xavier Savatier. 2017. A study of vicon system positioning performance. Sensors 17, 7 (2017), 1591.

[26] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. 2018. Learning to estimate 3D human pose and shape from a single color image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 459–468.

[27] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 652–660.

[28] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Advances in neural information processing systems. 5099–5108.

[29] Sandeep Rao. 2017. Introduction to mmWave sensing: FMCW radars. Texas Instruments (TI) mmWave Training Series (2017).

[30] Arindam Sengupta, Feng Jin, and Siyang Cao. 2020. NLP based Skeletal Pose Estimation using mmWave Radar Point-Cloud: A Simulation Approach. In 2020 IEEE Radar Conference (RadarConf20). IEEE, 1–6.

[31] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. 2020. mm-Pose: Real-Time Human Skeletal Posture Estimation using mmWave Radars and CNNs. IEEE Sensors Journal (2020).

[32] Karly A Smith, Clément Csech, David Murdoch, and George Shaker. 2018. Gesture recognition using mm-wave sensor for human-car interface. IEEE Sensors Letters 2, 2 (2018), 1–4.

[33] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. 2019. Human mesh recovery from monocular images via a skeleton-disentangled representation. In Proceedings of the IEEE International Conference on Computer Vision. 5349–5358.

[34] Fei Wang, Stanislav Panev, Ziyi Dai, Jinsong Han, and Dong Huang. 2019. Can WiFi estimate person pose? arXiv preprint arXiv:1904.00277 (2019).

[35] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. 2019. Person-in-WiFi: Fine-grained person perception using WiFi. In Proceedings of the IEEE International Conference on Computer Vision. 5452–5461.

[36] Jialu Wang, Zongqing Lu, and Qingmin Liao. 2019. Estimating Human Shape Under Clothing from Single Frontal View Point Cloud of a Dressed Human. In 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 330–334.

[37] Kangkan Wang, Jin Xie, Guofeng Zhang, Lei Liu, and Jian Yang. 2020. Sequential 3D Human Pose and Shape Estimation From Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7275–7284.

[38] Yong Wang, Wen Wang, Mu Zhou, Aihu Ren, and Zengshan Tian. 2020. Remote Monitoring of Human Vital Signs Based on 77-GHz mm-Wave FMCW Radar. Sensors 20, 10 (2020), 2999.

[39] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services. 14–26.

[40] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. Monoperfcap: Human performance capture from monocular video. ACM Transactions on Graphics (ToG) 37, 2 (2018), 1–15.

[41] Xin Yang, Jian Liu, Yingying Chen, Xiaonan Guo, and Yucheng Xie. 2020. MU-ID: Multi-user Identification Through Gaits Using Millimeter Wave Radios. In IEEE INFOCOM 2020-IEEE Conference on Computer Communications. IEEE, 2589–2598.

[42] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. 2017. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In Proceedings of the IEEE International Conference on Computer Vision. 910–919.

[43] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. 2017. Deep sets. In Advances in neural information processing systems. 3391–3401.

[44] Yunze Zeng, Parth H Pathak, Zhicheng Yang, and Prasant Mohapatra. 2016. Human tracking and activity monitoring using 60 GHz mmWave. In 2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN). IEEE, 1–2.

[45] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. 2019. Predicting 3d human dynamics from video. In Proceedings of the IEEE International Conference on Computer Vision. 7114–7123.

[46] Renyuan Zhang and Siyang Cao. 2018. Real-time human motion behavior detection via CNN using mmWave radar. IEEE Sensors Letters 3, 2 (2018), 1–4.

[47] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7356–7365.

[48] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, and Dina Katabi. 2019. Through-wall human mesh recovery using radio signals. In Proceedings of the IEEE International Conference on Computer Vision. 10113–10122.

[49] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D skeletons. In Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication. 267–281.

[50] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5745–5753.

[51] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chi Xu, Minglun Gong, and Li Cheng. 2020. 3D Human Shape Reconstruction from a Polarization Image. arXiv preprint arXiv:2007.09268 (2020).