M^{4}esh: mmWave-based 3D Human Mesh Construction for Multiple Subjects

Hongfei Xue[1†], Qiming Cao[2†], Yan Ju[1], Haochen Hu[1], Haoyu Wang[2], Aidong Zhang[3], Lu Su[2*]
[1]State University of New York at Buffalo, Buffalo, NY, USA [2]Purdue University, West Lafayette, IN, USA
[3]The University of Virginia, Charlottesville, VA, USA
Email: [1]{hongfeix,yanju,haochenh}@buffalo.edu, [2]{cao393,wang5346,lusu}@purdue.edu, [3]aidong@virginia.edu

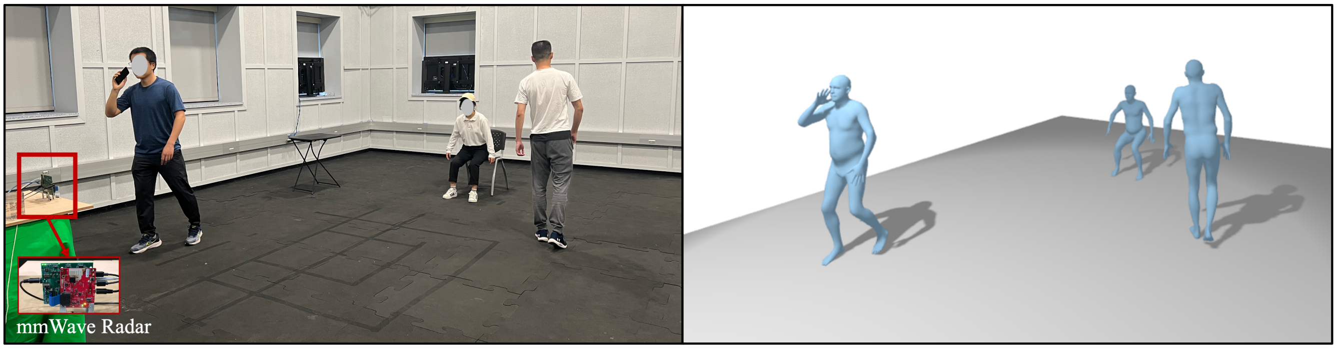


Figure 1: Illustration of mmWave-based 3D Human Mesh Construction for Multiple Subjects.

ABSTRACT

The recent proliferation of various wireless sensing systems and applications demonstrates the advantages of radio frequency (RF) signals over traditional camera-based solutions that are faced with various challenges, such as occlusions and poor lighting conditions. Towards the ultimate goal of imaging human body using RF signals, researchers have been exploring the possibility of constructing the human mesh, a structure capturing not only the pose but also the shape of the human body, from RF signals. In this paper, we introduce M^{4}esh, a novel system that utilizes commercial millimeter wave (mmWave) radar for multi-subject 3D human mesh construction. Our M^{4}esh system can detect and track the subjects on a 2D energy map by predicting the subject bounding boxes on the map, and tackle the subjects' mutual occlusion through utilizing the location, velocity and size information of the subjects' bounding boxes from the previous frames as a clue to estimate the bounding box in the current frame. Through extensive experiments on a real-world COTS millimeter-wave testbed, we show that our proposed M^{4}esh system can accurately localize the subjects and generate their human meshes, which demonstrate the superior effectiveness of the proposed M^{4}esh system.

CCS CONCEPTS

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

KEYWORDS

Wireless Sensing, Multiple Subjects, Human Mesh Estimation, Point Cloud, Millimeter Wave, Deep Learning

ACM Reference Format:
Hongfei Xue[1†], Qiming Cao[2†], Yan Ju[1], Haochen Hu[1], Haoyu Wang[2], Aidong Zhang[3], Lu Su[2]. 2022. M^{4}esh: mmWave-based 3D Human Mesh Construction for Multiple Subjects. In *The 20th ACM Conference on Embedded Networked Sensor Systems (SenSys '22), November 6–9, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3560905.3568545

1 INTRODUCTION

In recent years, significant efforts have been put toward building intelligent wireless sensing systems, with the goal of perceiving and understanding human activities using pervasive wireless signals. Thus far, various wireless sensing systems and applications have been proposed, demonstrating the advantages of RF signals over traditional camera-based solutions that are faced with various challenges, such as occlusions and poor lighting conditions.

Towards the ultimate goal of imaging human body using RF signals, the pioneer work, RF-Avatar [86], demonstrated that RF signals contain sufficient information to construct the human mesh, a structure capturing not only the pose but also the shape of the human body. However, the testbed [1] of RF-Avatar is built based on a specialized RF device consisting of a carefully assembled bulky antenna array and expensive USPRs, which limits its popularity and applications in the real world.

[*] Lu Su is the corresponding author.
[†] The first two authors contributed equally to this work.

SenSys '22, November 6–9, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9886-2/22/11…$15.00
https://doi.org/10.1145/3560905.3568545

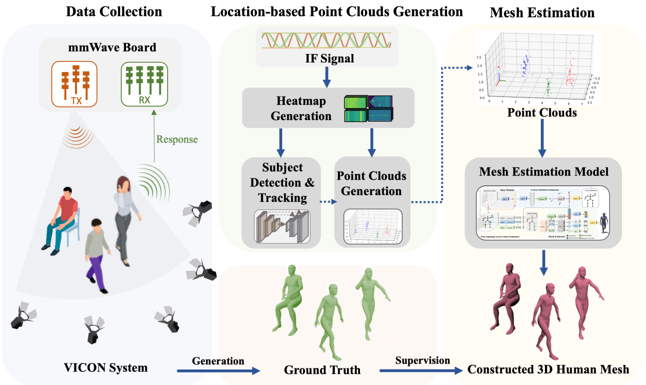


Figure 2: System Overview

Another seminal work, mmMesh [75], addresses this challenge and uses a commercial mmWave radar to achieve accurate human mesh estimation by transforming mmWave signals into 3D point clouds. However, mmMesh can only construct the mesh of a single subject. Two deficiencies in the design of mmMesh model restrict its extendibility to more pervasive multi-subject scenarios. First, due to the absence of a subject detection module, mmMesh is not capable of estimating the number of subjects. Second, the point cloud generation module of mmMesh may generate unevenly distributed points on multiple subjects, where the subjects far away from the radar may have insufficient points.

In order to conduct multi-subject human mesh estimation using mmWave radar, two challenges need to be addressed. The first challenge is to detect and locate all the subjects in the scenario where the subject number is usually unknown *a priori*. The second challenge is the mutual occlusion among multi-subjects when they conduct activities in proximity to each other. Since the mmWave signal cannot traverse the human body, short-term mutual occlusions among subjects are inevitable. These short-term occlusions damage the fluency of the constructed meshes.

To address these challenges, we propose **M^{4}esh**, an **MM**Wave-based 3D **M**ulti-Subject **Mesh** Construction System. In the design of M^{4}esh, our model is able to detect and track the subjects on a 2D energy map by predicting the subject bounding boxes on the map. To tackle the challenge of mutual occlusion, the proposed M^{4}esh system is capable of utilizing the location, velocity, and size information of the subjects' bounding boxes from the previous frames as a clue to estimate the bounding box in the current frame, even if the subjects are occluded by others. Additionally, we propose a coarse-to-fine mesh estimation strategy by carefully studying the detailed local structure of the point cloud based on a coarsely estimated human skeleton and further refining each joint representation by learning the relations among the joints based on the skeletal topology.

In order to evaluate the proposed M^{4}esh framework, we build a real-world testbed of our system using COTS millimeter wave devices. The evaluation results show that our proposed M^{4}esh system can accurately localize the subjects and generate the corresponding human meshes, which demonstrates the superior effectiveness of the proposed M^{4}esh system.

2 SYSTEM OVERVIEW

In this paper, we consider a real-world scenario where several human subjects are monitored by a mmWave radar whose emitted signals are reflected back by both the human bodies and ambient objects. Our proposed M^{4}esh system is able to adaptively generate the point clouds from the reflected signals of each subject based on his/her location, and estimate accurate human meshes by carefully studying the local structure around each joint and learning the relations among joints using the skeletal topology. As illustrated in Figure 2, our proposed M^{4}esh system is mainly composed of three components: data collection, location-based point clouds generation, and mesh construction.

2.1 Data Collection

This component is designed to collect the mmWave signals and corresponding ground truth human meshes when multiple subjects conduct activities surrounding the mmWave radar. In this process, the commercial mmWave radar keeps mixing the reflected signals with its emitting FMCW signals into the IF (Intermediate Frequency) signals. The IF signal will be utilized to generate point clouds in the next step. Meanwhile, the VICON motion capture system records high-precision dynamic pose information of these subjects which can be leveraged to generate the ground truth human mesh.

2.2 Location-based Point Clouds Generation

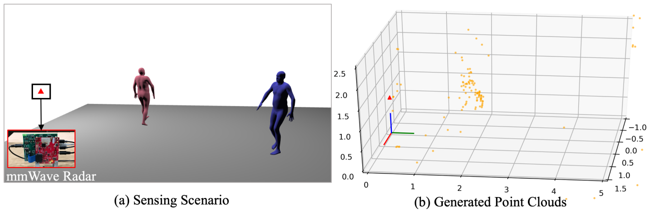
This component is designed to generate the point cloud for each subject as the input to the mesh estimation model. In our design, instead of generating the point cloud from the whole energy map, we generate the point clouds near each subject based on the location of the subject. Specifically, we first generated the required energy map from the IF signal, including Range-FFT heatmap, Doppler-FFT heatmap and MVDR-based 2D locational energy map. Then, we conduct subject detection and tracking on the 2D locational energy map using a deep detector. Last, based on the obtained locations of the subjects on the heatmap, we generate the point clouds for each subject.

2.3 Mesh Construction

The goal of this component is to construct the human mesh from the point cloud of each subject. In this component, we propose a novel deep learning model that can estimate accurate 3D human mesh by using a coarse-to-fine strategy to carefully learn the local structure information of the point cloud. Meanwhile, we also incorporate the topology of the human skeleton as prior knowledge in our model to learn the relations among the subject joints.

3 METHODOLOGY

To estimate 3D real-time human mesh of multiple subjects using the commercial mmWave radar, we first generate the point cloud for each subject based on his/her location (section 3.1). Then, based on the obtained point cloud for each subject, we estimate the human mesh (section 3.2). In Section 3.3, we briefly summarize the designs in our proposed framework to address the problem of short-term subject occlusion.



(a) Sensing Scenario

(b) Generated Point Clouds

Figure 3: Point Clouds Generation using Method in [75]

3.1 Location-based Point Clouds Generation

To tackle the task of human mesh estimation using commercial mmWave radar, the point cloud generated from mmWave signals has been proved as an effective input for the mesh construction models [75]. In [75], the pixels with top energy values (i.e., reflective signal strength) on the whole Doppler-FFT energy map will are transformed into the points in 3D space. The point clouds are generated based on these pixels by estimating their azimuths and elevations. However, this point cloud generation method cannot be directly applied in the multiple-subject scenarios. As we know, the strength of the reflected signal is usually higher when the subject is closer to the radar. When there are multiple subjects surrounding the mmWave radar, the strength of the reflective signals from the subjects who are close to the radar would dominate on the Doppler-FFT energy map. Correspondingly, there will be very few points generated from the reflective signal of the distant subjects.

In practice, the number of subjects is usually unknown *a priori*. In this case, the human mesh estimation model may miss the subjects far away from the radar. As illustrated in Figure 3 (a), there are two subjects walking in front of the mmWave radar (annotated as the red triangle) where the subject with blue mesh is far away from the radar and the subject with red mesh is close to the radar. And the point clouds generated by the method in [75] are shown as orange points in Figure 3 (b). We can see that the distribution of the generated points is biased to the subject close to the radar (with red mesh). Without sufficient points generated, the mesh estimation for the subject with blue mesh cannot be accurate (it may not be even detected in some cases). This situation is inevitable in the task of multiple subject mesh estimation since the distances from the subjects to the radar vary a lot. What's even worse is that the generated point cloud are very sparse (i.e., dozens of points) due to the low resolution of the commercial mmWave radar. The problem of the biased point distribution is exacerbated. Thus, how to generate sufficient number of points on the subjects who have different distances to the mmWave radar is a challenging problem.

A naive solution is to transform more pixels on the Doppler-FFT energy map into the 3D points despite their low-energy values. However, these generated 3D points with low-energy values are likely to have inaccurate estimations of locations or even to be fake points. As a result, the points with high-energy values will be overwhelmed, and hence the performance of the mesh estimation model is impaired.

To tackle this problem, we propose a location-based point cloud generation method. Instead of generating the point cloud from the whole energy map, our location-based point cloud generation method first localizes the subjects and then selects the pixels with top energy values only in the proximity of each subject. In this way, we can generate enough points for every subject, while reducing the noisy points from the ambience. The result of our method is illustrated in Figure 4 (g). Compared with the result of the method from [75], our method not only generates sufficient number of points on the subject far away from the mmWave radar, but also suppresses the noise from the ambient objects.

The steps of the proposed location-based point cloud generation framework are illustrated in the Figure 4. We first generate the Range-FFT heatmap (b) from the IF signal (a). Then, we generate a 2D locational energy map (c) using the MVDR [9] (Minimum Variance Distortionless Response) algorithm which assigns higher energy value to the subjects/objects with stronger reflective signal power in different ranges and angles (detailed in Section 3.1.1). Next, a subject detection and tracking deep model is applied on the locational energy map, and predicts the bounding box for each subject (e) (detailed in Section 3.1.2). Note that, the locational energy map generated based on the MVDR algorithm cannot be used to generate the point cloud, since the phase information among different antennas is missing after applying the MVDR algorithm, which is needed to estimate the elevation of each point. Instead, based on the predicted bounding boxes of the subjects, we crop the Doppler-FFT heatmap (d) into patches (f), each of which is cropped based on the range scope of the corresponding bounding box (detailed in Section 3.1.3). Then, the point cloud generation method in [75] is utilized on each patch to generate the intermediate point clouds (h) related to a subject. Lastly, the intermediate points are filtered out if their x-axis values are out of the x-axis scope of the corresponding bounding boxes, and obtain the final point cloud (g).

3.1.1 MVDR-based Locational Energy Map Generation. In order to localize the subjects, our first step is to generate a 2D locational energy map as the input of the detection model. In our design, we utilize the MVDR [9] (Minimum Variance Distortionless Response) algorithm, a data-adaptive beamforming solution, to generate the 2D locational energy map. The reasons are twofold. First, the MVDR algorithm is able to estimate the energy power from any direction by analyzing the information from the antenna array, which makes it ideal to infer the locations of the moving subjects. Second, in mmWave sensing scenarios, the noise and the desired mmWave signals are usually uncorrelated. Based on this principle, the MVDR algorithm is able to mitigate the effect of the noise and hence makes the locations of the subject on the energy map prominent.

We calculate the 2D locational energy map by applying the MVDR algorithm to calculate the reflective power of the signal from different angles on different range slots. Specifically, as illustrated in the Figure 4, we use a matrix $A \in \mathcal{R}^{N_p \times N_c \times N_s}$ to denote the input IF signal, where N_p is the number of transceiver pairs, N_c is the number of FMCW chirps in one frame and N_s is the number of ADC sampling points in each chirp. We first perform the Range-FFT operation (i.e., FFT along the ADC sampling axis), and obtained the matrix $B \in \mathcal{R}^{N_p \times N_c \times N_s}$. Then, we apply MVDR algorithm to calculate the energy distribution in different range bins and angles to calculate the locational energy map as follows. We use (x, r) to denote the locations where x is its value on the x axis and r is its distance to the mmWave transceivers (we assume the 1D radar array is arranged on the x axis). Then, the range bin index of the location (x, r) can be denoted as $\tilde{r} = \lfloor \frac{r}{\Delta r} \rfloor$ where we use Δr to denote the span of a range bin (i.e., the range resolution of the

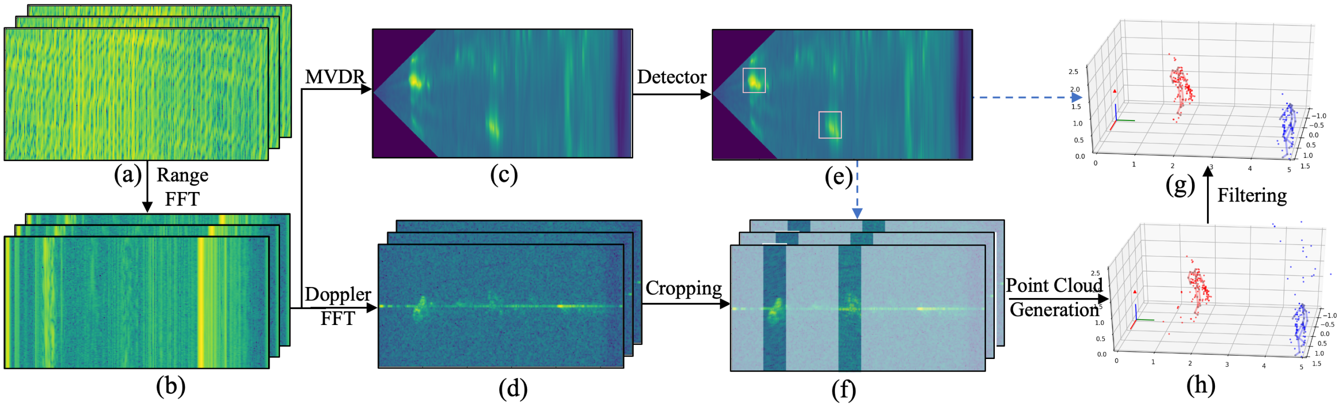


Figure 4: Point Clouds Generation. (a) IF signal; (b) Range-FFT Heatmaps; (c) 2D Locational Energy Map; (d) Doppler-FFT Heatmaps; (e) Predicted Bounding Box of the Subjects; (f) Cropped Local Heatmap Patches; (g) Final Point Clouds; (h) Intermediate Point Clouds

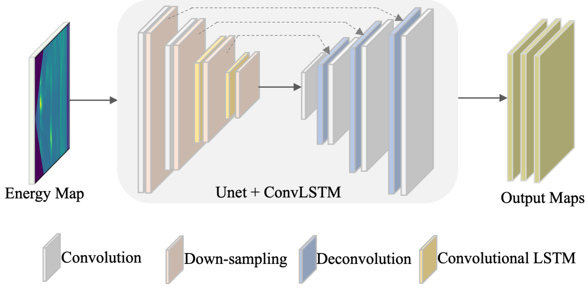


Figure 5: Detection and Tracking Network

mmWave radar). Thus, the corresponding signal data of location (x, r) is the $\tilde{r}$-th slicing of the matrix B along the ADC sampling axis (i.e., $B_{\tilde{r}}$). Thus, the signal power on the location (x, r) can be denoted as:

$$P(x,r) = \frac{1}{S^{H}_{(x,r)} \cdot (B_{\tilde{r}} B_{\tilde{r}}{}^{H})^{-1} \cdot S_{(x,r)}},$$

where H denotes the conjugate transpose of the data, and $S_{(x,r)}$ denotes the steering vector from the angle of the location (x, r). If we use $\theta_{(x,r)}$ to denote the angle in the steering vector, in our scenario, we have:

$$\begin{aligned} S_{(x,r)} &= [1, exp(-j\pi sin\theta_{(x,r)}), ..., exp(-j\pi(N_p - 1)sin\theta_{(x,r)})]^T, \\ &= [1, exp(-j\pi\frac{x}{r}), ..., exp(-j\pi(N_p - 1)\frac{x}{r})]^T, \end{aligned}$$

where T is the transpose of the vector. By calculating the power on all the desired locations, we obtain the locational energy map $C \in \mathcal{R}^{N_x \times N_s}$ as illustrated in the Figure 4 (c), where N_x is the predefined number of bins on the x-axis. Note that the locations where $|x| > |r|$ have no values in the locational energy map (i.e., two triangular areas in dark blue of Figure 4 (c)).

3.1.2 Subjects Detection and Tracking. Existing localization methods only consider the subject as a single point. However, to generate the points around the subject, we need to first specify the area of the subject via a bounding box. In this part, we will introduce how our proposed deep learning model detects and tracks the subjects using bounding boxes on the generated 2D locational energy map. As illustrated in Figure 5, our model takes the locational energy map as input and leverages U-Net [53] as the backbone structure to predict output maps. U-Net is a commonly used network structure in the task of image-to-image translation. It has an encoder and a decoder with skip connections between mirrored layers in the encoder and decoder stacks. However, to best fit the U-Net structure in our problem setting, which needs a recurrent structure to handle the energy map data along the time axis, we use ConvLSTM [56] layers to replace the last several convolutional layers of U-Net encoder in our design.

For the output, we employ the output format of the CenterTrack [92], which predicts the center maps, the size maps, the velocity maps, and the visibility maps of the bounding boxes of all subjects, simultaneously. To specify, each subject $e_i = (\mathbf{p}, \mathbf{s}, \mathbf{v}, vis)$ in the 2D locational energy map is represented by its center $\mathbf{p} \in \mathcal{R}^2$, size $\mathbf{s} \in \mathcal{R}^2$, velocity $\mathbf{v} \in \mathcal{R}^2$ and visibility $vis \in \mathcal{R}$, where center is the bounding box center, size is the height and width of the bounding box, velocity is the distance from the bounding box center in the current frame to the corresponding bounding box center in the next frame, and the visibility score indicates the percentage of the bounding box is occluded by the bounding boxes of other subjects from the view of the mmWave radar in the current frame.

Correspondingly, the deep model generates heatmaps to output these values. The produced center map $\hat{Y^c} \in [0,1]^{N_x \times N_s}$ indicates the location of the bounding box center, where a prediction $Y^{\hat{c}}_{hw} = 1$ corresponds to a detected bounding box center, while $Y^{\hat{c}}_{hw} = 0$ is the background point, where h and w are the indexes on along map height and width. For the corresponding ground truth bounding box center onto a heatmap $Y^c \in [0,1]^{N_x \times N_s}$, we use a Gaussian kernel $Y_{hw} = exp(-\frac{(h-\tilde{p}_h)^2+(w-\tilde{p}_w)^2}{2\sigma_p^2})$, where σ_p is an object size-adaptive standard deviation [31], $\tilde{p}_h$ and $\tilde{p}_w$ are the ground truth bounding box center indexes. Since there are multiple subjects, the Gaussians of the subjects overlap. We take the element-wise maximum [8] during the calculation of the ground truth map Y^c. The training objective is a penalty-reduced pixel-wise logistic regression with focal loss [33]:

$$L^p = \frac{-1}{N} \sum_{hw} \begin{cases} (1 - \hat{Y}_{hw})^{\alpha} log(\hat{Y}_{hw}) & \text{if } Y_{hw} = 1 \\ (1 - Y_{hw})^{\beta} (\hat{Y}_{hw})^{\alpha} log(1 - \hat{Y}_{hw}) & \text{Otherwise} \end{cases}$$

where α and β are hyper-parameters of the focal loss, and N is the number of the subjects in the energy map. For the produced size map $\hat{Y}^s \in R^{2 \times N_x \times N_s}$, we use L1 loss to regress the bounding box size:

$$L^s = \frac{1}{N} \sum_{k=1}^{N} |\hat{Y}^s_{p_i} - Y^s_{p_i}|$$

where Y^s is the ground truth size map, and the supervision acts only at the bounding box center locations p_i. Similarly, for the produced velocity map $\hat{Y}^v \in R^{2 \times N_x \times N_s}$ and the visibility map[63] $\hat{Y}^{vis} \in R^{N_x \times N_s}$, we also use L1 loss the regress them:

$$L^v = \frac{1}{N} \sum_{k=1}^{N} |\hat{Y}^v_{p_i} - Y^v_{p_i}|, \quad L^{vis} = \frac{1}{N} \sum_{k=1}^{N} |\hat{Y}^{vis}_{p_i} - Y^{vis}_{p_i}|,$$

where Y^v and Y^{vis} are the ground truth velocity map and the ground truth visibility map. The model loss is the summation of four losses:

$$Loss = \gamma^p * L^p + \gamma^s * L^s + \gamma^v * L^v + \gamma^{vis} * L^{vis}, \tag{1}$$

where γ^p, γ^s, γ^v and γ^{vis} are hyperparameters.

At inference time, we extract the peaks in the center map whose value is greater than a detection threshold 0.3 as well as all the other values of the pixels from its 15 local area. Let $\hat{p}_i$ denote the locations of the i-th bounding box center, then the size, the velocity and the visibility score of this bounding box are $\hat{Y}^s_{\hat{p}_i}$, $\hat{Y}^v_{\hat{p}_i}$ and $\hat{Y}^{vis}_{\hat{p}_i}$. As we can see in Figure 4 (e), the bounding boxes of two subjects on the locational energy map are detected.

To track the subject among different frames, for each detection at position $\hat{p}_t$ on the frame t, we greedily associate it with the closest position $\hat{p}_{t-1} + \hat{v}_{t-1}$ within the bounding box of $\hat{p}_t$, where $\hat{v}_{t-1}$ is the predicted speed of $\hat{p}_{t-1}$. Then, we can obtain the trajectory of each subject. If the closest position $\hat{p}_{t-1} + \hat{v}_{t-1}$ is not found within the bounding box of $\hat{p}_t$, and this subject is detected on the energy map for over 3 frames, we spawn a new tracklet for this new subject. If there is no position $\hat{p}_t$ found which can be associated with the position $\hat{p}_{t-1} + \hat{v}_{t-1}$, we hallucinate the position $\hat{p}_{t-1} + \hat{v}_{t-1}$. If the associated positions are not found for 7 frames, we end this tracklet.

3.1.3 Point Clouds Generation. Having the bounding box on the locational energy map, we can generate point clouds based on the predicted bounding boxes of the subjects. Note that, the locational energy map generated based on the MVDR algorithm cannot be used to generate the point cloud. This is because the phase information among different antennas, which is needed to estimate the elevation of each point, will be missing after the MVDR algorithm is employed. Instead, we use the range scope information of the bounding boxes to extract local patches from Doppler-FFT energy maps (Figure 4 (d)) for the generation of the point cloud. Also, we leverage the x-axis scope information of the bounding boxes to filter out the points which are far away from the subjects.

Specifically, as illustrated in Figure 4 (d), we first conduct Doppler-FFT on the matrix B from section 3.1.1. Then, we extract the local patch based on the range scope of each bounding box as shown in Figure 4 (f). Based on the extracted local patch, we extract the pixels with top energy values, and estimate their elevation and azimuth based on their phase among different antennas [75]. The generated point cloud is shown in Figure 4 (h). The points in red are generated based on the bounding box of the red skeleton, while The points in blue are generated based on the bounding box of the blue skeleton. Lastly, based on the x-axis scope of the bounding box, we filter out the points whose values on the x-axis are out of the corresponding x-axis scope as shown in Figure 4 (g). Compared with the holistic point cloud generation method in Figure 3 (b), our proposed location-based point cloud generation method not only produces sufficient points for each subject, but also suppresses the noise points from the ambient objects. What's more, since the points are generated based on the subjects' bounding box, we don't need to use another algorithm to assign point clouds to different subjects.

3.2 Coarse-to-Fine Human Mesh Estimation

After obtaining the point cloud for each subject, we then feed the point cloud data of each subject into the proposed mesh estimation model to predict the subject's mesh. However, the estimation result may be inaccurate if we simply infer the mesh from the holistic points without considering the local structure of the point cloud. Specifically, to better estimate the human mesh from the point clouds, the local structure of the point clouds near the joints of the subject should be precisely captured. However, it is difficult to directly tell which point is associated with which part of the human body. To address this problem, we propose a coarse-to-fine mesh estimation framework where a coarse human skeleton is predicted using the whole point clouds. Then, based on the joint locations of the coarse skeleton, the local structure of the point cloud near each joint is learned by utilizing attention-based point feature aggregation to predict a fine-grained human mesh. As we know, the local structure of the point cloud is highly related to the current pose of the subject, since a point cloud represents the set of reflective points on a subject. Learning the point cloud structure based on coarsely estimated pose of the subject can help the model exploit the local information. In addition, we also notice that the location and angle of adjacent joints are highly correlated. Thus, we proposed to incorporate the human skeletal topology information as prior knowledge for accurate mesh estimation by modeling the human skeleton as a graph.

As illustrated in Figure 6, the mesh estimation model is basically composed of four modules. The base module is used to extract features from each point, the Coarse Skeleton Estimator aims to estimate a coarse skeleton of the subject based on the information of holistic points, the pose-aware Joint Estimator refines the joint representations by carefully studying the local structure around each joint and learning the relations among joints using the skeletal topology (graph), and Mesh Estimator reconstructs the human mesh by taking the joint representation from the Pose-aware Joint Estimator and the global representation from the Coarse Skeleton Estimator as input.

3.2.1 Base Module. As illustrated in Figure 6, the input of this module consists of the feature vectors of all the 3D points in the point cloud, including the x,y,z coordinates, the range value, the radius velocity value and the energy value (on Doppler-FFT energy map) of the points. The feature vectors of all the points are stacked into a 2D matrix as input for the Base Module. Then, the output of this module is obtained by feeding each feature vector into a share-weighted MLP [48, 49] (Multi-Layer Perceptron) to generate a higher level representation for each point.

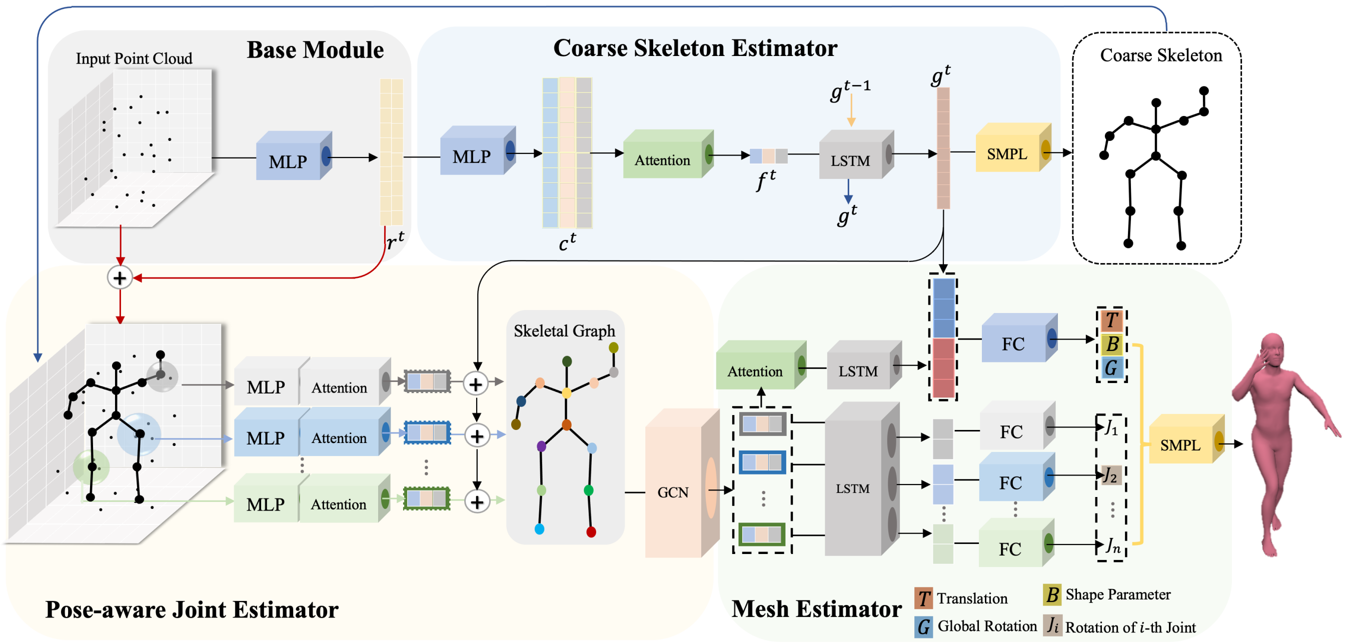


Figure 6: Mesh Estimation Model

3.2.2 *Coarse Skeleton Estimator.* In this part, we coarsely estimate the skeleton of the subject based on the information of the holistic point cloud. Similar to the feature extraction operation in the Base Module, we also use a MLP to extract higher-level feature representation for each point. Then, we use the Attention block to aggregate the representations of all points in the current frame:

$$f^t = \sum_{i \in N^t} L(c_i^t) \cdot c_i^t, \tag{2}$$

where t is the index of the current frame, f^t is the aggregated representation, c_i^t is the representation of each point, N^t is the number of points in current frame and L is a linear mapping function. The function of the attention block is the weighted sum of all point features without information loss and it allows us to dynamically learn relative contribution of each point [20]. It is worth noting that the attention operation is invariant to point permutation, and it can be used to aggregate 3D point cloud features [79]. Then, a three-layer LSTM [18] block is applied to generate the global representation g^t.

To generate the human skeleton, we leverage the 3D human body model SMPL [38] (Skinned Multi-Person Linear model) in our paper. SMPL is a widely used parametric human model that estimates both 3D human mesh and skeleton 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 deformation 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 and 24 human skeleton joints by taking 10 shape parameters and 72 pose parameters as input. By feeding the global representation g^t into the SMPL module, we obtained the coarse skeleton of the subject.

3.2.3 *Pose-aware Joint Estimator.* After obtaining the coarse skeleton of the subject, in this part, we aim to learn a more accurate human pose and shape representation by learning the local structure of the point cloud based on the coarsely estimated pose of the subject and incorporating the topological structure of the human skeleton. Traditionally, to learn the local structures of the point cloud, a sampling method is first used to sample some points from the point cloud as grouping center. Then, the points are grouped into several subsets based on the grouping center. Finally, the representations of all the points in each subset are aggregated together as the local structure representation [75].

In our design, since we have coarsely estimated the subject's skeleton, we can regard the joints of the skeleton as the grouping centers. As we know, the point clouds are generated when the subject is conducting the pose as estimated by the coarse skeleton. Associating the human pose and the distribution of the point cloud helps the network to better extract the information from the point cloud. What's more, using the joints as grouping centers can help the model analyze the local point structure around each joint. Thus, it helps the model learn more accurate joint representations. Then we feed each group of the local point clouds, which are concatenated with their high-level representation obtained from the Base Module, into the share-weighted MLP block and the Attention blocks to generate the pose-aware representation for each joint. We concatenate the global representation from the Coarse Skeleton Estimator with each pose-aware joint representation to encode the relationship among local joints and the global skeleton as the input of the next step.

Then, the proposed model learns the inherent topological structure of the human skeleton by utilizing the GCN on the skeletal graph as illustrated in the Figure 6. GCN [26] is a special variant of Convolutional Graph Neural Networks (ConvGNNs). It has demonstrated powerful ability to learn node representations by jointly encoding graph structures and node features by generalizing the operation of convolution from grid data to graph data [83]. To specify, for the skeletal graph $G(V, E)$ where V is the set of nodes (i.e., joints) and E is the set of edges (i.e., the connections between joints by limbs or torso). For each node v_i, $e_{ij} = (v_i, v_j) \in E$ denotes an edge pointing from v_j to v_i and its neighborhood is defined as $N(v_i) = \{u \in V | (v_i, u) \in E\}$. The adjacency matrix A is a $n \times n$ matrix with $A_{ij} = 1$ if $e_{ij} \in E$ and $A_{ij} = 0$ if $e_{ij} \notin E$. The node attributes of the skeleton graph is denoted as X, where $X \in R^{n\times d}$ is a node feature matrix with $x_v \in \mathcal{R}^d$ representing the node representation vector of a node v. The main idea of GCN is to generate a node v's representation by aggregating its own features x_v and neighbors' features x_u, where $u \in N(v)$. It defines a simplified convolution layer as:

$$H = f(\overline{A}X\Theta) \tag{3}$$

where $\overline{A} = \widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}$. $\widetilde{A} = A + I$ denotes the adjacency matrix with inserted self-loops and $\widetilde{D}$ is the diagonal degree matrix of $\widetilde{A}$ with $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$. Θ is a learn-able parameter matrix and f is an activation function. In our model design, we apply the GCN three times on the skeletal graph and obtained the pose-aware joint representations.

3.2.4 Mesh Estimator. In this part, we use the global information to predict the global rotations, translations and shape parameters of the subject, and use the local joint information to predict the pose of the subject. Specifically, to obtain the final global representation, we use another Attention block to fuse all the pose-aware joint representations and feed the aggregated representation into an LSTM block. Then we concatenate the obtained vector and the global representation from the Coarse Skeleton Estimator. We used the concatenated vector to predict the global information of the subject, including the translation, the shape and the global rotation of the subject using an FC block (i.e., linear mapping). For each joint rotation of the subject, we first feed each pose-aware joint representation into an LSTM block, respectively. Then, the output of the LSTM is fed into the FC block to generate 6D representation [93] to represent the rotation of each joint. After obtaining both the global parameter and the joint parameter of the subject, we can feed the parameters into the SMPL module to generate the final human mesh.

3.2.5 Model Loss. The model loss of this pose-aware human mesh estimator is the summation of six components as follows:

$$Loss = \sum_{K \in \{V,S,B,V_c,S_c,B_c,\}} \alpha_K * \sum_{t}^{T} \|K^t - \mathcal{GT}(K^t)\|_{L_1} \tag{4}$$

Here we use V,S,B to denote the vertex matrix, the skeleton matrix and the shape matrix obtained from the Pose-aware Mesh Estimator. Similarly, V_c, S_c, B_c denote the vertex matrix, the skeleton matrix and the shape matrix obtained from the Coarse Skeleton Estimator. We use $GT(K)$ to denote the corresponding ground truth of the generated matrix K. α_K denotes the hyper-parameters.

3.3 Summary of Designs to Address the Short-term Subject Occlusion Problem

In this section, we briefly summarize the designs in our proposed framework to address the problem of short-term subject occlusion, where the distant subject may be occluded by the other subjects in the multiple subject mesh estimation scenarios:

(1) We add a recurrent structure in the model design in both the detection and tracking model and the mesh estimation model. The recurrent structure allows the model to conduct inference based on the prior knowledge from previous frames, even when the subjects are occluded by other subjects. In the detection and tracking model, the bounding boxes can be hallucinated with the support of the ConvLSTM in the occluded scenario. And in the mesh estimation model, we leverage LSTM blocks in both Coarse Skeleton Estimator and Mesh Estimator to tackle the occlusion problem.

(2) We force the detection and tracking model to predict the visibility for the bounding box of each subject, where visibility is a score to indicate the percentage of the bounding box is occluded by the bounding boxes of other subjects from the view of the mmWave radar in the current frame. With the supervision of the visibility score, the deep learning model can learn the geometric relationships among the subjects.

(3) The detection and tracking model outputs the velocity of each bounding box in addition to its location and size. When the model fails to predict of the bounding box when the subject is occluded even with the aforementioned first and the second schemes, we use the velocity and the bounding box position from the previous frame to infer the positions of the missing bounding boxes in the current frame.

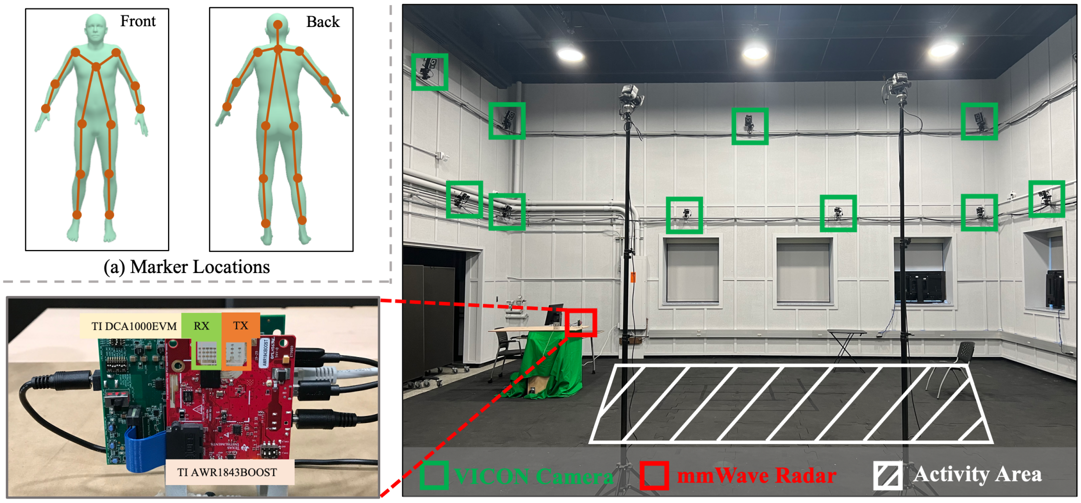
(4) The SMPL block in our mesh estimation model guarantees to produce a complete human mesh when the subject is partly occluded or even fully occluded, as long as the bounding box of the subject is detected.

(5) During the training process, we feed the ground truth label to the models no matter whether the subject is occluded or not. In this way, potentially, the detection and tracking model learns to correctly predict the bounding box and the mesh estimation model learns to accurately generate the mesh even when the subject is occluded.

4 EXPERIMENTS

4.1 Testbeds

4.1.1 VICON System. We use VICON system [65] to collect high-precision 3D human pose ground truth for model training. As shown in Figure 7 (c), the VICON system consists of 21 VICON Vantage cameras that can emit and receive infrared light. The emitted infrared light will be reflected by the markers covered with highly reflective materials, and then be captured by the VICON Vantage cameras. Totally, 27 high-precision pearl makers are placed on each human subject near his/her joints, so that the VICON acquires the positions of the joints and generates precise human skeletons based on the locations of markers. Figure 7 (a) shows the positions of these markers on a subject. According to [42], the errors caused by the location of each marker is less than 2 millimeters which is



(a) Marker Locations

(b) mmWave Radar

(c) Experimental Scenario

Figure 7: The Testbeds and the Experimental Scenario

accurate enough in our scenario. Besides, the sampling rate of the VICON system is set to 10 frames per second.

4.1.2 mmWave Testbed. As illustrated in Figure 7 (b), we use TI AWR1843BOOST mmWave radar [61] to collect mmWave data and an additional TI DCA1000EVM data-capture adapter to stream the data from the mmWave radar. The mmWave radar consists of 3 transmitting antennas and 4 receiving antennas. The transmitting antennas successively emit FMCW chirps, and the receiving antennas capture the signals reflected by the human bodies and ambient objects. In our experiments, we set the bandwidth of the transmitted FMCW chirp to 3.9GHZ, which ranges from 77GHZ to 80.9GHZ. Here each frame is composed of 128 chirps, and each chirp is composed of 256 sampling points. Based on our device setting, our mmWave device can reach up to 11 m sensing range, 4.3 cm range resolution, 4.5 m/s maximum sensing velocity and 7.1 cm/s velocity resolution. Besides, to align with the setting of the VICON system, the mmWave radar is also set to send 10 frames per second. In the experiment, we place the mmWave testbed on a table whose height is 92 cm, and the distance between the mmWave testbed and the activity area is about 1.5 m, where the activity area is a rectangle with a length of 4.6 m and a width of 3.0 m.

4.2 Data Collection and Prepossessing

4.2.1 Data Collection. We invite 18 volunteers to perform multi-person daily activities within the activity area that can be captured by the VICON system. To simulate real-world scenarios, the participants can freely conduct and change their activities from the following activities: walking in circles, walking back and forth in straight, picking up the phone from the desk, putting down the phone on the desk, answering phone calls while walking, playing with the cell phone while sitting on the chair, sitting on the chair and standing up from the chair. The activities will be performed by either two or three volunteers simultaneously. The total time length of the data is 510 mins.

4.2.2 Ground Truth Mesh Construction. We use Skinned Multi-Person Linear Model (SMPL) model to produce the ground truth human mesh for model training. Note that we utilize the neutral SMPL model in our paper, where the gender information of the subject is no longer required. The SMPL model takes the pose information, shape information, and translation information of the subject as input and outputs 3D human skeleton as well as 3D human mesh. The pose and translation information can be directly obtained from the VICON system. Specifically, We calculate the 3D position of human joints by averaging the locations of two corresponding markers which are stuck on the back and front of the subject near the corresponding joints. Note that the calculated VICON joints and SMPL pose vectors are not one-on-one mappings, the SMPL model requires more pose vectors than that can be calculated from our setting. Since the missing joints have little effect on the designed daily activities, we simply set those pose vectors to constant values. The ground truth shape vector can be obtained using the approach in [3]. To best match the human mesh model with the ground truth height of the subject, we also manually adjust the shape vector values. Note that the ground truth bounding box on the 2D locational energy map is generated based on the bounding box of the ground truth human mesh.

4.3 Model Settings and Model Training

We describe the implementation details for both the detection and tracking model and the mesh estimation model in this section. For the detection and tracking model, we employ 6 down-sampling layers in the encoder, where the sampling rate is $\frac{1}{2}$. Before the first three down-sampling layers, we add 2 layers of convolution operation, and before the last three down-sampling layers, we add 1 convolutional layer and 1 ConvLSTM layer. Correspondingly, there are also 6 deconvolutional layers, whose up-sampling rate is 2. And after each deconvolutional layer, we employ 2 layers of convolution operation. During the model training, the batch size is set to 16 and the sequence length in each batch is set to 16. The parameters α and β in the focal loss are set to 2 and 4, respectively. The parameter γ^p, γ^s, γ^v and γ^{vis} in equation 1 are set to 1, 5, 1 and 5, respectively. For the mesh estimation model, all the MPL blocks and the LSTM blocks, GCN blocks have 3 layers. The number of grouped points around each joint in the Pose-aware Joint Estimator is set to 16. During the model training, the hyperparameter α_K is set to 0.001 for V and V_c, 1 for S and S_c, 0.1 for B and B_c. The initial learning rates for both models are set to 0.001, which decays during the training process. In the model training, we also randomly move the ground truth bounding boxes for a small distance in each batch during the point cloud generation step to train a robust mesh estimator. Besides, we use PyTorch [45] to implement both deep learning models and train the model on NVIDIA A6000 GPU. For basic settings, we train both models with the first 2200 frames of the data and use the rest 500 frames for evaluation. Additionally, we also conducted cross-trial experiments in Section 4.5.6.

4.4 Baselines and Metrics

4.4.1 Baselines. In this paper, we compare our model with the following baselines:

RF-Avatar [86] is one of the pioneering works to reconstruct human mesh using wireless signals. The model proposed in RF-Avatar takes as input 3D RF energy map from a specialized RF device [1] in each frame and generates human meshes in two stages. In the first stage, it detects and tracks each subject using the Trajectory Proposal Network (TPN) based on MaskRCNN [17], and crops the located regions in the energy map. The second stage of the model in RF-Avatar receives the cropped features and uses a Trajectory-CNN (TCNN) to produce the SMPL parameters. However, different from the layout of antennas in [1] where both the transmitting and the receiving antennas are arranged on a T-shaped holder, the receiving antennas of the commercial mmWave radar are arranged in a straight line. Hence, we can only generate 2D energy maps from the commercial mmWave radar, and modify the input interface of RF-Avatar correspondingly for 2D energy maps.

mmMesh [75] has demonstrated its superior performance in the task of single-person mesh reconstruction using the point clouds generated by commercial mmWave radar. In this baseline, we feed the point cloud for each subject generated by our model as input to the original mmMesh model.

Holistic PC. This baseline uses exactly the same models in our framework except that the input point clouds are generated using the method in [75]. The generated point clouds will be filtered based on the ground truth bounding box of each subject.

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

Average Vertex Error (V) [7, 86]. We compute the average vertex error by averaging the Euclidean distance between all 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 performance of the location error, pose error, shape error, and gender error.

Average Joint Localization Error (S) [22, 86]. 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.

Procrustes-Aligned Mean Per Joint Position Error (PA-S) [27]. In this metric, we first align the predicted 3D poses to the ground truth poses using Procrustes [13], where the errors from both the translation and the orientation are eliminated. Then, the average Euclidean distance between the joint locations of the predicted human mesh and the ground truths for all the subjects and activities is measured.

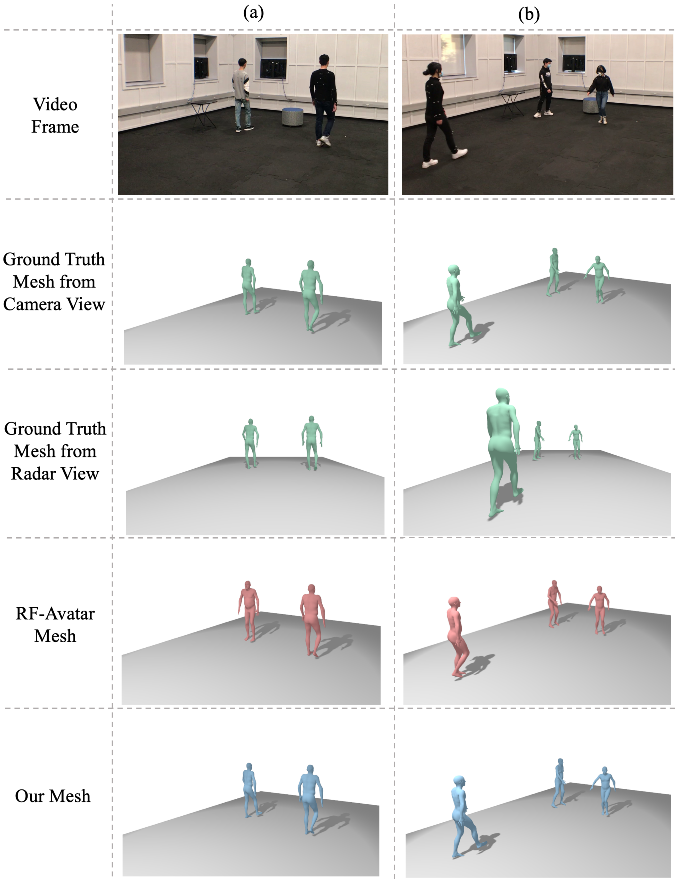

(a)
(b)
Video Frame
Ground Truth Mesh from Camera View
Ground Truth Mesh from Radar View
RF-Avatar Mesh
Our Mesh

Figure 8: Results for 2 and 3 Subjects Scenarios without Occlusion. (a) 2 Subjects (b) 3 Subjects

Average Joint Rotation Error (Q) [75]. 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 ground truth rotations. In this metric, we only consider the rotations of shoulder joints, elbow joints, hip joints, and knee joints from both sides of the subject, since these are the most relevant joints to the human pose.

Mesh Localization Error (T) [75]. 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.

4.5 Experiment Results

4.5.1 Qualitative Results for the Scenarios without Mutual Occlusion. In this part, we present the qualitative results when multiple subjects are conducting the activities without occluding each other. As illustrated in Figure 8, the first row of the figure presents the video frames where the subjects with different shapes are conducting activities. As we can see, in column (a), two subjects are

Model	V(cm)	S(cm)	PA-S(cm)	Q(°)	T(cm)
RF-Avatar	9.76	9.35	3.22	8.33	6.74
Holistic PC	4.53	4.31	1.96	4.66	3.17
mmMesh	4.54	4.32	1.93	4.50	3.21
M^{4}esh	**3.74**	**3.54**	**1.67**	**3.87**	**2.61**
M^{4}esh + GT Bbox	**3.72**	**3.52**	**1.66**	**3.86**	**2.60**

Table 1: Results for 2 Subjects Scenario without Occlusion

Model	V(cm)	S(cm)	PA-S(cm)	Q(°)	T(cm)
RF-Avatar	10.51	9.70	3.62	10.43	6.41
Holistic PC	6.83	6.24	2.62	6.58	4.64
mmMesh	6.51	5.90	2.71	6.72	4.49
M^{4}esh	**4.97**	**4.64**	**2.04**	**4.84**	**3.54**
M^{4}esh + GT Bbox	**4.90**	**4.50**	**2.02**	**4.91**	**3.42**

Table 2: Results for 3 Subjects Scenario without Occlusion

walking in the same direction. While, in column (b), there are two subjects walking in the opposite direction, and the third subject just puts down the phone on the table and starts to walk. The second and third rows show the ground truth meshes generated based on the VICON motion capture system. The only difference is that the meshes in the second row are rendered from the perspective of the camera in the first row, while the meshes in the third row are rendered from the perspective of the mmWave radar. From the mesh rendering results, we know that there are no occlusions among the subjects. The fourth row presents the results of the RF-Avatar model. As we can see, in column (a) RF-Avatar model has a wrong prediction for the orientation of the left subject. In column (b), the poses of three estimated meshes by RF-Avatar model are not accurate. One possible reason is that the information contained in the 2D energy map may be ambiguous and inaccurate, since the information of one dimension is missing on the 2D energy map. The last row shows the results of our proposed M^{4}esh. We can see that the meshes produced by our model look very similar to the ground truth meshes. This is because our location-based point cloud generation method can generate sound and complete point clouds for each subject, which provides sufficient information to the mesh estimation model. What's more, the coarse-to-fine mesh estimation strategy also helps the model generate more accurate meshes.

4.5.2 Quantitative Results for Scenarios without Mutual Occlusion. In this section, we quantitatively evaluate the performance of our proposed model compared with the baselines based on the metrics described in section 4.4.2. As shown in Table 1 and Table 2, M^{4}esh outperforms all baselines in all of the metrics which demonstrates the effectiveness of our proposed model to produce accurate human meshes in the multiple subject scenarios.

To prove the effectiveness of our proposed location-based point cloud generation method, we can compare our results with the baseline Holistic PC. As we can see, the mesh estimation performance increases when we use the proposed location-based point cloud generation method to address the biased point cloud distribution problem in the multi-subject human mesh estimation task.

As illustrated in the two tables, our proposed model outperforms the baseline mmMesh. It demonstrates the effectiveness of our design in the mesh estimation model where our model carefully studies the local structure of the point cloud based on a coarsely estimated human skeleton and incorporates the skeletal information by treating the skeleton structure as a graph.

Comparing the results of our proposed model and RF-Avatar, we achieve much better mesh estimation results. There are mainly two reasons. First, we design the Pose-aware Joint Estimator in our model, where the joint representations are carefully refined by learning the local point structure around each joint and encoding the relationship among subject joints. However, RF-Avatar outputs all the joint representations at one time without refining them or considering the relationships among joints. Second, our model takes the point cloud with 3D information as input, while the input of RF-Avatar is 2D energy map. Originally, RF-Avatar takes 3D energy map from a specialized RF device [1] as input. However, we can only obtain 2D energy map due to the antenna layout on the commercial mmWave radar. Additionally, as we can see in the two tables, the average vertex error only increases within 0.1 cm when we use the bounding boxes predicted by our model. This demonstrates the effectiveness of our detection and tracking network and the robustness of our mesh estimation model.

Comparing the results in Table 1 and Table 2, we can find that the performances of all the models decrease when the number of subjects is increased from two to three. The reasons behind are twofold. First, the accuracy of predicted bounding boxes from the subject localization models drops when the subject number increases, since the locational relationships among three subjects are more complicated than that among two subjects. And the increased localization errors are passed to the mesh estimation stage. Second, the multi-path effect among the subjects and the ambient objects is aggravated when there are more subjects in the scenario, which results in the lower quality of the generated point clouds and the heatmap. Hence, the performance of the mesh estimation model is impaired. Although the performance drop is inevitable when the subject number increases, our model has the least decrease (1.2 cm) among all the models which demonstrates the robustness of our model when the number of subjects varies.

4.5.3 Qualitative Results for the Scenarios with Mutual Occlusion. In this section, we consider the challenging scenarios where the subjects are occluded by each other from the view of the mmWave. As illustrated in Figure 9, the first three rows show the video frames, ground truth meshes from the camera view, and the ground truth meshes from the mmWave radar view when subjects are conducting daily activities. As we can see, in column (a), two subjects are walking in front of the radar, while one subject is completely occluded by the other subject. In column (b), two subjects are walking and the third subject is sitting on a chair and making a phone call. In this scenario, one walking subject is also occluded by the other walking subject. The last three rows show ground truth bounding boxes, the bounding boxes generated by RF-Avatar, and the bounding boxes generated by our proposed model on the 2D locational energy map. As we can see, the RF-Avatar algorithm fails to generate the bounding boxes of the occluded subjects, while our model successfully estimates the bounding boxes with accurate locations and sizes for

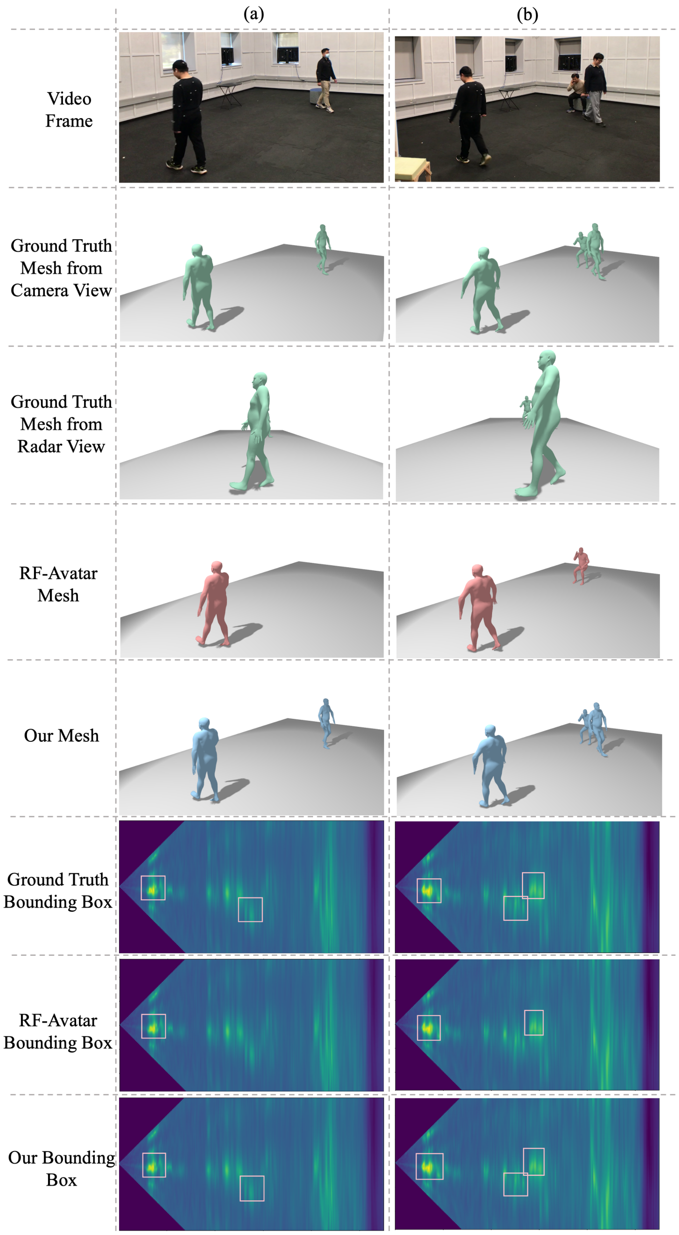

(a)
(b)
Video Frame
Ground Truth Mesh from Camera View
Ground Truth Mesh from Radar View
RF-Avatar Mesh
Our Mesh
Ground Truth Bounding Box
RF-Avatar Bounding Box
Our Bounding Box

Figure 9: Results for 2 and 3 Subjects Scenarios with Occlusion. (a) 2 Subjects (b) 3 Subjects

Model	V(cm)	S(cm)	PA-S(cm)	Q(°)	T(cm)
RF-Avatar	10.27	9.76	2.52	9.59	6.29
mmMesh	6.24	6.01	2.07	5.10	4.94
M^{4}esh	**4.77**	**4.48**	**1.93**	**4.49**	**3.42**
M^{4}esh + GT Bbox	**4.74**	**4.45**	**1.90**	**4.42**	**3.45**

Table 3: Results for 2 Subjects Scenario with Occlusion

Model	V(cm)	S(cm)	PA-S(cm)	Q(°)	T(cm)
RF-Avatar	11.57	10.71	4.07	10.69	7.73
mmMesh	7.78	7.12	2.85	8.32	5.64
M^{4}esh	**5.64**	**5.10**	**2.64**	**6.96**	**3.61**
M^{4}esh + GT Bbox	**5.32**	**4.48**	**2.51**	**6.56**	**3.36**

Table 4: Results for 3 Subjects Scenario with Occlusion

Model	2-wo	3-wo	2-w	3-w
RF-Avatar (AP@70)	0.921	0.859	0.868	0.854
M^{4}esh (AP@70)	**0.960**	**0.925**	**0.916**	**0.865**
RF-Avatar (AP)	0.631	0.609	0.593	0.565
M^{4}esh (AP)	**0.681**	**0.650**	**0.630**	**0.586**

Table 5: Detection Results for Different Scenarios

all the subjects. This is because the reflective energy of the occluded subject is similar to the energy in the background, which is hard to be detected using the mask-rcnn-based algorithm in the RF-Avatar. In contrast, our model is capable to utilize the location, velocity, and size information of the bounding boxes from the previous frames as a clue to estimate the bounding box in the current frame, even if the reflective pattern of the subject on the energy map is not distinct. Correspondingly, the human meshes are predicted in the fourth row and the fifth row by the RF-Avatar model and our model, respectively. As we can see, RF-Avatar cannot construct the occluded subject in both scenarios, since the subject is not detected during the localization. Meanwhile, our mesh estimation model still predicts complete and accurate human meshes compared with the ground truth mesh. This is achieved by incorporating the SMPL block and the recurrent structure in the design of the mesh estimation model. To demonstrate the smoothness of the predicted human mesh results in the time dimension, we illustrate consecutive video frames as well as corresponding ground truth and predicted meshes in Figure 10 for 3 subject scenario with occlusion. To cover longer activity interval, we show a frame every 0.4s. In the figure, the three subjects are conducting 'putting down the phone', 'walking', and 'standing up from the stool', respectively. And the sitting subject is briefly occluded by the walking subject. By comparing our predicted results with the ground truth and video frames, we can see that our model is capable of constructing smooth and realistic human meshes for multiple subjects, as well as handling short-time occlusions among subjects.

4.5.4 Quantitative Results for Scenario with Mutual Occlusion. In this section, we quantitatively evaluate the results of our model in the scenarios with subject occlusions. As shown in Table 3 and Table 4, M^{4}esh can still achieve better results than all the baselines on all metrics, which demonstrates the robustness of our model even when there are occlusions among the subjects. Compared to the results in Table 1 and Table 2 where there is no occlusion among the subjects, we can see the performance of all the models has dropped. Since the subjects are occluded, there is little information about the occluded subjects. The mesh estimation models have to rely only on the historical pose and shape information to estimate the mesh in the current frame. Hence, the results of the model estimation can be very inaccurate in these scenarios.

4.5.5 Results of Bounding Box Prediction. As illustrated in table 5, we measure the performance of the bounding box prediction results of RF-Avatar and our proposed M^{4}esh model using both AP [34] and AP@70 (the higher the better). AP measures the performance

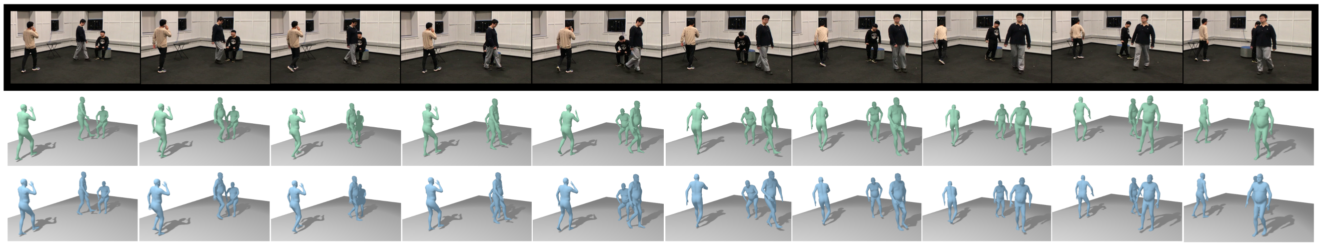


Figure 10: Consecutive Frames for 3 Subject Scenario with Occlusion The first row shows the video frames. The second and third rows show the ground truth meshes and the predicted meshes, respectively. We illustrate every fourth frame for a better illustration of the scenario.

Scenario	V(cm)	S(cm)	PA-S(cm)	Q(°)	T(cm)
Cross-trial (2-wo)	4.51	4.14	2.13	4.69	2.75
Cross-trial (3-w)	7.24	6.47	3.08	8.49	4.71
Cross-obstacle (2-wo)	7.28	6.85	2.96	7.09	5.45
Cross-obstacle (3-w)	9.50	8.76	3.65	9.28	7.29

Table 6: Results for Cross-trial and Cross-obstacle Settings

of the subject detection performance by considering both precision and recall rates between the predicted bounding boxes and the ground truth bounding boxes under certain Intersection over Union (IoU) thresholds. Specifically, it calculates the average area under the Precision-Recall Curve with the IoU threshold from 0.5 (i.e., AP@50) to 0.95 (i.e., AP@95) with a step size of 0.05. Especially, AP@70 is calculated with the IoU threshold at 0.7 where a predicted bounding box should have a large overlapping ratio with the ground truth bounding box to be considered as a correct prediction. As we can see in the table, our model outperforms the baseline for all the scenarios on both criteria, which demonstrates the effectiveness of the proposed detection and tracking network.

4.5.6 Results of the Cross-trial Experiment. We also conduct a cross-trial experiment where the subject combinations or the activities of subjects are different among different trials. In this experiment, the training data and the evaluation data are collected in different trials. We report the results of both 2-subject scenario without occlusion (the simplest scenario) and 3-subject scenario with occlusion (the most complex scenario) in the first and second rows of Table 6, respectively. By comparing them with the corresponding results in Table 1 and Table 4, we can find that the performances of our proposed model drop by small margins (0.77 cm and 1.60 cm), which demonstrate our model is robust enough for different training settings.

4.5.7 Results of Experiment with Blockage. To demonstrate the robustness of the proposed model, we use a bamboo panel as the obstacle between the radar and the subjects. The results are reported in the third and fourth rows of Table 6. Similarly, we report the results of both the 2-subject scenario without occlusion (the simplest scenario) and the 3-subject scenario with occlusion (the most complex scenario), respectively. By comparing them with the results in Table 1 and Table 4, the performance of the proposed model decreases when there are obstacles between the radar and subjects. This is because the signal patterns are affected by the blockage.

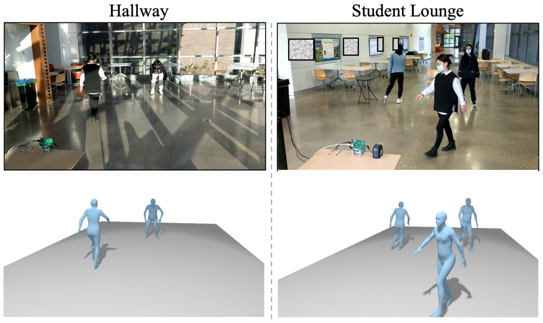


Figure 11: Cross-environment Experiments

4.5.8 Results of Cross-environment Experiment. To study the robustness of our proposed model in unseen environments, we conducted human mesh inference experiments in the hallway and the student lounge, which are different from the environment where the training data are collected. The qualitative results are illustrated in figure 11. As we can see in the figure, our model correctly estimates the meshes of the subjects in different environments, which proves the robustness of the proposed M^{4}esh in new environments.

4.5.9 System Complexity and Latency. In this section, we analyze the computational complexity and the latency of the proposed system. For the complexity of our proposed M^{4}esh model, we report the number of trainable parameters in our deep model. The overall number of trainable parameters is $2.29M$, where our detection and tracking module has 0.39M parameters and our mesh estimator has 1.90M parameters. We can see that our proposed M^{4}esh is a lightweight mesh estimation model, compared with the baseline RF-Avatar (53M parameters) and VIBE (a state-of-the-art real-time mesh estimation model in computer vision which has 48M parameters). For the latency of the proposed system, we conducted the inference on a desktop (NVIDIA A6000 GPU, Intel Xeon Gold 6254 CPU), and report the average latency. The whole inference process is divided into six steps: collection of radar data (including data capturing and transferring), calculation of heatmap, localization of subjects (including the bounding box prediction using velocity information), generation of point clouds, estimation of human meshes, visualization of human meshes. Their latencies are 115.0ms, 22.9ms, 6.0ms, 8.9ms, 6.0ms and 43.3ms. In total, the inference latency of the whole system is 202.1ms, which demonstrates our proposed system can work in real-time.

4.5.10 Limitations. Although we have utilized several designs to address the short-term subject occlusion problem, the proposed model still has some limitations. For example, our model can only handle short-term occlusion, where the bounding box of the occluded subject can be estimated within certain frames after occlusion. We also assume the occluded subject keeps performing the same activity and moving in a straight line, which is not always true. When the subject changes their directions or activities during the occlusion, the model prediction can be incorrect.

5 RELATED WORK

mmWave-based Localization: Compared with WiFi-based and acoustic-based localization techniques, mmWave signal could improve the localization resolution to millimeter-level thanks to its small wavelength [70]. Generally, mmWave-based localization needs two parameters that are range and Angle of Arrival (AOA). Range can be obtained from Channel Impulse Response (CIR) [51, 71], phase changes [70], and frequency changes with FMCW [2, 75, 84, 90]. AoAs are usually estimated by beamforming techniques such as MUSIC and MVDR [35, 43, 51, 71]. Beam steering [70] could also be utilized to estimate the AoA of mmWave signals. Joint estimation of Received Signal Strength (RSS) and AoA can also enable 3D indoor localization in [35]. All above mmWave-based localization techniques consider the human body as a single point. In this paper, to determine where to generate the point cloud for each subject, we need to infer the area of each subject in space. It requires the model to enclose the human body by a bounding box instead of treating it as a single point. Thus, in this paper, we modify the object detection algorithm in computer vision and apply it to the generated 2D locational energy map to detect and localize the subjects by bounding boxes. Recently, there are also works that apply point cloud techniques to track human bodies [2, 90]. However, in these works, cloud points are utilized to calculate the centroid to represent the position of human body.

mmWave-based Sensing: Millimeter Wave (mmWave) radar has recently become popular for versatile sensing applications including vibration measurement [15, 16, 19], motion estimation and capture [4, 40, 89], imaging [82], Fruit Quality Assessment [77], indoor mapping [39], material sensing [72] and voice related sensing [32, 37, 94]. In the domain of human sensing, mmWave radar is also applied to conduct human tracking and identification/authentication [36, 51, 76, 90, 91], human activity recognition [30], vital sign monitoring [25, 67, 69, 73, 88], subject detection [57]. Different from these applications, our work focuses on the task of human mesh estimation using mmWave radar.

Human Pose Estimation from Wireless Signals: In recent years, many wireless sensing systems have been developed to estimate human pose [1, 22, 52, 54, 55, 66, 68, 85, 87]. Among them, [66, 68, 85] focus on 2D pose estimation, and [22, 29, 54, 55, 87] are capable of estimating 3D human pose. Different from these works, our work aims to estimate 3D human mesh, which contains both pose and shape information of the subjects.

3D Human Mesh Construction from Wireless Signals: As far as we know, there are limited works that have explored 3D human mesh construction task using wireless signals. RF-Avatar [86] is a pioneering work that demonstrates the RF signals contain sufficient information for the human mesh estimation. What's more, [86] also justifies the effectiveness of the RF signal to tackle the technical challenges faced by the traditional camera-based solutions, such as occlusions and poor lighting conditions. However, the testbed [1] of RF-Avatar is built based on the specialized RF device consisting of a carefully assembled bulky antenna array and expensive USPRs, which limits its popularity and real-world applications. In contrast, our proposed M^{4}esh utilizes a commercial portable mmWave radar and achieves superior performance on the task of multiple subject human mesh estimation. mmMesh [75] is another seminal work that uses the commercial mmWave radar to achieve accurate single-subject human mesh estimations. The success of the mmMesh demonstrates the great potential of commercial mmWave devices on the fine-grained human sensing tasks. However, the framework of mmMesh cannot be directly applied in the multiple subject scenario for two reasons. First, due to the absence of a subject detection module in the design of mmMesh framework, mmMesh is not capable of estimating the number of the subjects which is required before the mesh estimation step in the multi-subject scenario. Second, the point cloud generation method in the mmMesh is not adaptable in the multi-subject scenario, since the biased point clouds can be generated where the distant subjects may have no sufficient numbers of points for mesh estimations as detailed in 3.1.

3D Human Mesh Estimation in Computer Vision: Since the release of statistical body models such as SCAPE [5] and SMPL [38], 3D human mesh recovery has been drawing broader attention [62]. Substantial methods have been proposed for the task of human mesh estimation based on images [11, 12, 14, 21, 23, 28, 44, 47, 58, 59, 74, 80, 81] and videos [6, 10, 24, 27, 41, 46, 50, 60, 64, 78]. Despite the great success achieved by image/video based approaches, the performance of these methods can be severely impaired by bad illumination, occlusion, and blur. More importantly, privacy issues occur when cameras are deployed to monitor the human subjects. In contrast, our mmWave-signal-based solutions are immune to poor lighting and occlusion conditions, as well as, alleviating the privacy issue.

6 CONCLUSIONS

In this paper, we investigate the possibility of using commercial mmWave radar to conduct 3D human mesh construction for multiple subjects. Specifically, we propose M^{4}esh, a novel mmWave-based sensing system that can detect and track the subjects on a 2D energy map by predicting the bounding boxes of the subjects on the map, and tackle the subjects' mutual occlusion through utilizing the location, velocity and size information of the subjects' bounding boxes from the previous frames as a clue to estimate the bounding box in the current frame, even if the subjects are occluded by others. Additionally, we propose a coarse-to-fine mesh estimation strategy by carefully studying the detailed local structure of the point cloud based on a coarsely estimated human skeleton and further refining each joint representation by learning the relations among the joints based on the skeletal topology. The accurate subject localization and mesh estimation results based on a real-world commercial mmWave sensing testbed demonstrate the effectiveness of the proposed M^{4}esh system.

ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation under Grants CNS-1652503, CNS-2213700, and CCF-2217071.

REFERENCES

[1] 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.

[2] Mohammad Arif Ul Alam, Md Mahmudur Rahman, and Jared Q Widberg. 2021. Palmar: Towards adaptive multi-inhabitant activity recognition in point-cloud technology. In IEEE INFOCOM 2021-IEEE Conference on Computer Communications. IEEE, 1–10.

[3] 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.

[4] Yasin Almalioglu, Mehmet Turan, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. 2020. Milli-RIO: Ego-motion estimation with low-cost millimetre-wave radar. IEEE Sensors Journal 21, 3 (2020), 3314–3323.

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. Scape: shape completion and animation of people. In ACM SIGGRAPH 2005 Papers. 408–416.

[6] Anurag Arnab, Carl Doersch, and Andrew Zisserman. 2019. Exploiting temporal context for 3D human pose estimation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3395–3404.

[7] 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.

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7291–7299.

[9] Jack Capon. 1969. High-resolution frequency-wavenumber spectrum analysis. Proc. IEEE 57, 8 (1969), 1408–1418.

[10] Carl Doersch and Andrew Zisserman. 2019. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. Advances in Neural Information Processing Systems 32 (2019).

[11] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. 2020. Three-dimensional reconstruction of human interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7214–7223.

[12] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. 2020. Hierarchical kinematic human mesh recovery. In European Conference on Computer Vision. Springer, 768–784.

[13] John C Gower. 1975. Generalized procrustes analysis. Psychometrika 40, 1 (1975), 33–51.

[14] Riza Alp Guler and Iasonas Kokkinos. 2019. Holopose: Holistic 3d human reconstruction in-the-wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10884–10894.

[15] Junchen Guo, Yuan He, Chengkun Jiang, Meng Jin, Shuai Li, Jia Zhang, Rui Xi, and Yunhao Liu. 2021. Measuring micrometer-level vibrations with mmwave radar. IEEE Transactions on Mobile Computing (2021).

[16] Junchen Guo, Meng Jin, Yuan He, Weiguo Wang, and Yunhao Liu. 2021. Dancing Waltz with Ghosts: Measuring Sub-mm-Level 2D Rotor Orbit with a Single mmWave Radar. In Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021). 77–92.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision. 2961–2969.

[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.

[19] 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.

[20] 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.

[21] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. 2020. Coherent reconstruction of multiple humans from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5579–5588.

[22] 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.

[23] 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.

[24] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. 2019. Learning 3d human dynamics from video. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 5614–5623.

[25] Abdelwahed Khamis, Branislav Kusy, Chun Tung Chou, Mary-Louise McLaws, and Wen Hu. 2020. RFWash: a weakly supervised tracking of hand hygiene technique. In Proceedings of the 18th Conference on Embedded Networked Sensor Systems. 572–584.

[26] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016).

[27] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. 2020. Vibe: Video inference for human body pose and shape estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 5253–5263.

[28] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2252–2261.

[29] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. m3Track: mmwave-based multi-user 3D posture tracking. In Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services. 491–503.

[30] 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.

[31] Hei Law and Jia Deng. 2018. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European conference on computer vision (ECCV). 734–750.

[32] 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.

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision. 2980–2988.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision. Springer, 740–755.

[35] Zhipeng Lin, Tiejun Lv, and P Takis Mathiopoulos. 2018. 3-D indoor positioning for millimeter-wave massive MIMO systems. IEEE Transactions on Communications 66, 6 (2018), 2472–2486.

[36] Jianwei Liu, Chaowei Xiao, Kaiyan Cui, Jinsong Han, Xian Xu, Kui Ren, and Xufei Mao. 2021. A Behavior Privacy Preserving Method towards RF Sensing. In 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS). IEEE, 1–10.

[37] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. In Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems. 97–110.

[38] 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.

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

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

[41] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2020. XNect: Real-time multi-person 3D motion capture with a single RGB camera. Acm Transactions On Graphics (TOG) 39, 4 (2020), 82–1.

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

[43] Alain Olivier, Guillermo Bielsa, Irene Tejado, Michele Zorzi, Joerg Widmer, and Paolo Casari. 2016. Lightweight indoor localization for 60-GHz millimeter wave systems. In 2016 13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON). IEEE, 1–9.

[44] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In 2018 international conference on 3D vision (3DV). IEEE, 484–494.

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[46] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. 2022. Human mesh recovery from multiple shots. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1485–1495.

[47] 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.

[48] 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.

[49] 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.

[50] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. 2021. Tracking people with 3D representations. arXiv preprint arXiv:2111.07868 (2021).

[51] Sai Deepika Regani, Chenshu Wu, Beibei Wang, Min Wu, and KJ Ray Liu. 2021. mmWrite: passive handwriting tracking using a single millimeter-wave radio. IEEE Internet of Things Journal 8, 17 (2021), 13291–13305.

[52] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2021. Winect: 3d human pose tracking for free-form activity using commodity wifi. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5, 4 (2021), 1–29.

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention. Springer, 234–241.

[54] 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 20, 17 (2020), 10032–10044.

[55] Cong Shi, Li Lu, Jian Liu, Yan Wang, Yingying Chen, and Jiadi Yu. 2022. mPose: Environment-and subject-agnostic 3D skeleton posture reconstruction leveraging a single mmWave device. Smart Health 23 (2022), 100228.

[56] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems 28 (2015).

[57] Xian Shuai, Yulin Shen, Yi Tang, Shuyao Shi, Luping Ji, and Guoliang Xing. 2021. millieye: A lightweight mmwave radar and camera fusion system for robust object detection. In Proceedings of the International Conference on Internet-of-Things Design and Implementation. 145–157.

[58] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. 2021. Monocular, one-stage, regression of multiple 3d people. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 11179–11188.

[59] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. 2022. Putting people in their place: Monocular regression of 3d people in depth. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13243–13252.

[60] 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/CVF International Conference on Computer Vision. 5349–5358.

[61] TI. 1930. Texas Instruments. http://www.ti.com

[62] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. 2022. Recovering 3d human mesh from monocular images: A survey. arXiv preprint arXiv:2203.01923 (2022).

[63] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. 20210. Learning to track with object permanence. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 10860–10869.

[64] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. 2017. Self-supervised learning of motion capture. Advances in Neural Information Processing Systems 30 (2017).

[65] VICON. 2008. VICON Motion Systems. https://www.vicon.com

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

[67] Fengyu Wang, Xiaolu Zeng, Chenshu Wu, Beibei Wang, and KJ Ray Liu. 2021. mmhrv: Contactless heart rate variability monitoring using millimeter-wave radio. IEEE Internet of Things Journal 8, 22 (2021), 16623–16636.

[68] 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.

[69] Weicheng Wang, Zhenhua Jia, Chenren Xu, Guojie Luo, Daqing Zhang, Ning An, and Yanyong Zhang. 2021. Feasibility study of practical vital sign detection using millimeter-wave radios. CCF Transactions on Pervasive Computing and Interaction 3, 4 (2021), 436–452.

[70] Teng Wei and Xinyu Zhang. 2015. mtrack: High-precision passive tracking using millimeter wave radios. In Proceedings of the 21st Annual International Conference on Mobile Computing and Networking. 117–129.

[71] Chenshu Wu, Feng Zhang, Beibei Wang, and KJ Ray Liu. 2020. mmTrack: Passive multi-person localization using commodity millimeter wave radio. In IEEE INFOCOM 2020-IEEE Conference on Computer Communications. IEEE, 2400–2409.

[72] Chenshu Wu, Feng Zhang, Beibei Wang, and KJ Ray Liu. 2020. mSense: Towards mobile material sensing with a single millimeter-wave radio. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4, 3 (2020), 1–20.

[73] Yucheng Xie, Ruizhe Jiang, Xiaonan Guo, Yan Wang, Jerry Cheng, and Yingying Chen. 2022. mmEat: Millimeter wave-enabled environment-invariant eating behavior monitoring. Smart Health 23 (2022), 100236.

[74] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. 2019. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 7760–7770.

[75] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. 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. 269–282.

[76] 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.

[77] Zhicheng Yang, Parth H Pathak, Mo Sha, Tingting Zhu, Junai Gan, Pengfei Hu, and Prasant Mohapatra. 2019. On the feasibility of estimating soluble sugar content using millimeter-wave. In Proceedings of the International Conference on Internet of Things Design and Implementation. 13–24.

[78] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. 2022. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11038–11049.

[79] 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.

[80] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In European Conference on Computer Vision. Springer, 465–481.

[81] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. 2018. Deep network for the integrated 3d sensing of multiple people in natural images. Advances in Neural Information Processing Systems 31 (2018).

[82] Diana Zhang, Akarsh Prabhakara, Sirajum Munir, Aswin Sankaranarayanan, and Swarun Kumar. 2021. A Hybrid mmWave and Camera System for Long-Range Depth Imaging. arXiv preprint arXiv:2106.07856 (2021).

[83] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. 2019. Graph convolutional networks: a comprehensive review. Computational Social Networks 6, 1 (2019), 1–23.

[84] Zhenyuan Zhang, Xiaojie Wang, Darong Huang, Xin Fang, Mu Zhou, and Ying Zhang. 2021. MRPT: Millimeter-Wave Radar-Based Pedestrian Trajectory Tracking for Autonomous Urban Driving. IEEE Transactions on Instrumentation and Measurement 71 (2021), 1–17.

[85] 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.

[86] 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.

[87] 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.

[88] Peijun Zhao, Chris Xiaoxuan Lu, Bing Wang, Changhao Chen, Linhai Xie, Mengyu Wang, Niki Trigoni, and Andrew Markham. 2020. Heart rate sensing with a robot mounted mmwave radar. In 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2812–2818.

[89] Peijun Zhao, Chris Xiaoxuan Lu, Bing Wang, Niki Trigoni, and Andrew Markham. 2021. 3D Motion Capture of an Unmodified Drone with Single-chip Millimeter

Wave Radar. In 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 5186–5192.

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

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

[92] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2020. Tracking objects as points. In European Conference on Computer Vision. Springer, 474–490.

[93] 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.

[94] Yinian Zhou, Awais Ahmad Siddiqi, Jia Zhang, Junchen Guo, Rui Xi, Meng Jin, Zhengang Zhai, and Yuan He. 2021. Voice Recovery from Human Surroundings with Millimeter Wave Radar. In IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 1–2.