SUPER: Seated Upper Body Pose Estimation using mmWave Radars

Bo Zhang, Zimeng Zhou, Boyu Jiang, Rong Zheng
Department of Computing and Software
McMaster University
Hamilton, ON, Canada
{*zhanb59, zhouz287, jiangb11, rzheng*}@mcmaster.ca

***Abstract*—In industrial countries, adults spend a considerable amount of time sedentary each day at work, driving and during activities of daily living. Characterizing the seated upper body human poses using mmWave radars is an important, yet under-studied topic with many applications in human-machine interaction, transportation and road safety. In this work, we devise SUPER, a framework for seated upper body human pose estimation that utilizes dual-mmWave radars in close proximity. A novel masking algorithm is proposed to coherently fuse data from the radars to generate intensity and Doppler point clouds with complementary information for high-motion but small radar cross section areas (e.g., upper extremities) and low-motion but large RCS areas (e.g. torso). A lightweight neural network extracts both global and local features of upper body and output pose parameters for the Skinned Multi-Person Linear (SMPL) model. Extensive leave-one-subject-out experiments on various motion sequences from multiple subjects show that SUPER outperforms a state-of-the-art baseline method by 30 – 184%. We also demonstrate its utility in a simple downstream task for hand-object interaction.**

***Index Terms*—Seated upper body pose estimation, mmWave radars, data fusion, point clouds, deep neural networks**

I. Introduction

Human pose estimation (HPE) estimates the configuration of human body parts from input data captured by sensors and has attracted much attention in industry and the research community due to its wide range of applications, including the human-machine interactions [1], fitness [2], virtual reality [3], smart home [4] and smart vehicle [5], etc. While full-body HPE is important in characterizing joint movements during locomotions, a 2019 study showed that adults ages 20 to 75 in the US reported spending an average of 9.5 hours sedentary each day [6]. Therefore, seated upper body human pose estimation (SUB-HPE) is arguably more relevant in interactive applications and understanding users' mental states (e.g., alertness and attention). For example, by monitoring upper limb movements while sitting, novel applications can be developed to empower users to control digital interfaces, manipulate augmented reality environments, and manage smart home systems. SUB-HPE can also find applications in transportation and road safety, where drowsy or inattentive drivers pose a significant risk on roadways. Analyzing head poses, hand placements and orientation of the upper body allows the detection of early signs of drowsiness or distraction.

In recent years, the rapid advancements in deep learning led to significant progress in human body modeling [7], [8] and HPE using various sensing modalities. Notable work in HPE includes OpenPose [9] and VitPose [10] in computer vision, Deep inertial poser [11] and IMUPoser [12] using IMU sensors, mmPose [13] and mmMesh [14] with mmWave radars, and DensePose [15] using WiFi devices, to name a few. Among different sensing modalities, mmWave radars offer distinct advantages due to their ability to penetrate obstructions like garments or walls, adapt to diverse lighting and weather conditions, and preserve user privacy. Furthermore, the substantial bandwidth (in the GHz range) equips mmWave radars with resilience against noise, interference, and center-meter level range resolutions. However, existing mmWave-based solutions predominantly target full-body locomotions and are not designed for handling nuanced upper-body movements. mmWave-based SUB-HPE shares with full-body HPE common challenges stemming from low spatial resolutions as the result of few on-board transmitting and receiving antennas on low-end commercial-of-the-shelf (COTS) mmWave radars, specular reflections and variations from inherent micro-body movements. But, crucially, it must also handle limited mobility in the upper body's core area when sitting, as well as the small radar cross-sections (RCS) of upper extremities, ranging from -45 dBsm to -20 dBsm for hands [16].

In this work, we devise SUPER, a framework for Seated Upper Body Pose Estimation using mmWave Radars. The framework encompasses a dual-radar pre-processing and fusion pipeline and a light weight neural network to predict upper body pose parameters. To increase the spatial resolution of the acquired radar data, two closely positioned radar sensors, oriented perpendicular to each other, are utilized. A novel dual-radar masking algorithm coherently fuses data from the radars to generate two complementary types of point clouds: the intensity point cloud (IPC) and the Doppler point cloud (DPC). The latter captures motion information of extremities while the former better characterizes low-motion portions of the upper body (e.g., torso areas). Benefiting from the sparse point cloud representation, the lightweight neural network extracts both global and local features of the upper body. Finally, the Skinned Multi-Person Linear (SMPL) model is applied to yield realistic human body poses and motions. An example of the data captured by an RGB camera, a motion capture system,

979-8-3503-7025-6/24/$31.00 ©2024 IEEE
DOI 10.1109/IoTDI61053.2024.00020

and the predicted and ground truth poses can be found in Figure 1.

We have implemented a prototype of SUPER utilizing two Texas Instruments IWR6843ISK mmWave radars[1]. A diverse group of 10 subjects, encompassing different genders, ages, and body mass indices (BMIs), were recruited for data collection in a laboratory setting. The data collection process involved subjects engaging in predefined arm, head, torso motion sequences. Experiment results show that SUPER consistently outperforms a state-of-the-art (SOTA) baseline method and achieves 112mm in average Mean Per Joint Position Error (MPJPE) and 15.89mm Procrustes alignment MPJPE (PA-MPJPE) metrics in leave-one-subject out trials. To demonstrate the utility of SUPER, we also implement and evaluate a simple downstream task of hand-object interaction.

In summary, we make the following new contributions toward mmWave-based fine-grained SUB-HPE in this work.

- In this work, we investigate a new task, i.e., SUB-HPE, and collect a dataset consisting of various head, torso as well as arm motions using mmWave radars.
- The proposed framework, SUPER, utilizes the intensity information from multi-antenna radar systems, to characterize the spatial occupation of human body under low mobility and Doppler information to capture motions of extremities.
- We demonstrate the feasibility of deploying two asynchronous but closely located mmWave radars to improve spatial resolution. A novel masking algorithm is proposed to coherently fuse data from both radars.
- SUPER has been evaluated using different motion sequences and data from a diverse set of users and shows superior performance compared to a SOTA baseline method.

The rest of the paper is organized as follows. A review of recent development of mmWave-based HPE methods and public datasets is presented in Section II. In Section III, we introduce the proposed pipeline and key techniques. Section IV provides experiment setups and the dataset we build. Detailed results and system performance are provided in Section V. Section VI demonstrates the potentials of the proposed system by a downstream task. Finally, we discuss the limitations of the work and conclude this paper in Section VII.

II. Related work

FMCW radars as an emerging technology have attracted significant attention and have been investigated in a variety of sensing tasks, e.g. tracking and localization [17]–[19], gesture recognition [20]–[23], and vital sign monitoring [24]–[27], etc. In this section, we focus on mmWave-based HPE methods and public datasets.

A. MmWave-based human pose estimation

In [13], Sengupta *et al.* present mm-Pose, which is among the first works in mmWave-based full-body HPE. mm-Pose projects radar point clouds from two separate and perpendicularly oriented radars onto the depth-azimuth(XY) and depth-elevation(XZ) plane, respectively to create two 2D intensity images. The images are then fed into a forked CNN structure to predict the human skeletal joints. In [28], An *et al.* propose the MARS system which takes 5D radar point clouds (x, y, z, intensity and Doppler) as input and outputs human pose in several rehabilitation scenarios. Xue *et al.* [14] introduce mmMesh which adopts PointNet [29] as the feature extractor of the point cloud and incorporates SMPL [7] to this task, facilitating both body shape and pose predictions. In a follow-up work to [14], multi-subject 3D human mesh construction is investigated [30]. This is achieved by obtaining the location information from an energy map, and selectively generating 4D point clouds close to the subjects. A fine-grained human mesh is then predicted using a coarse-to-fine mesh estimation framework. Most recently, instead of using radar point clouds, Lee *et al.* [31] introduce the velocity-specific range-doppler-azimuth-elevation map (VRDAEMap) as the input and developed a cross-modality training framework that fuses multi-scale radar features using a Cross- and Self-Attention Module (CSAM), and further refines the predicted key points through a Pose Refinement Graph Convolutional Networks (PRGCN).

The aforementioned works on mmWave-based HPE differ in the number of devices used for data collection, data representation (point clouds vs. images), and the backbone neural network architecture. However, none considers SUB-HPE, where there is typically limited trunk and lower limb mobility. A summary of the key aspects of these methods can be found in Table I.

B. Public mmWave-based HPE datasets

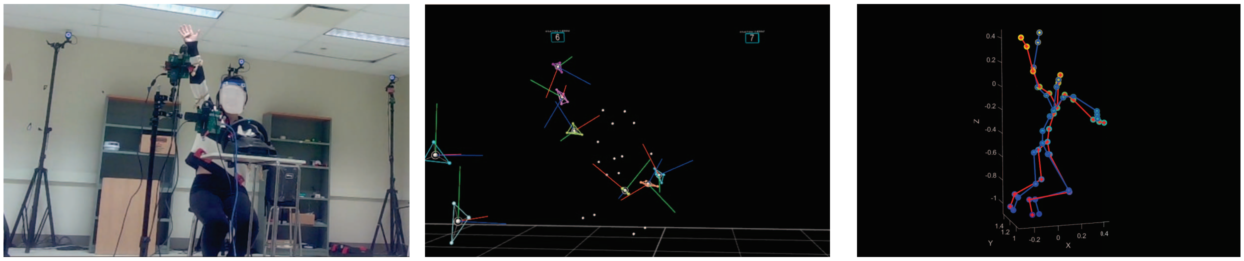
Very few public datasets are currently available for mmWave-based HPE. In [28], the authors release a dataset MARS containing radar point clouds and annotation obtained using Microsoft Kinect V2 sensor. Chen *et al.* proposed mmBody [32], a multi-scenario RGBD-paired mmWave radar (Arbe Robotics Phoenix) point cloud dataset for human pose reconstruction with 3D ground truth provided by a motion capture system. The work in [31] also provides a dataset HuPR, which contains raw radar data together with 3D annotation generated from a synced RGB camera.

With the exception of HuPR, the aforementioned public datasets only contain intermediate representations of the radar data, e.g., point clouds. The lack of raw data greatly limits innovations on radar signal processing algorithms and consequently affects the informativeness of training data to the HPE models. Another limitation of some datasets (e.g., HuPR and MARS) lies in the absence of accurate ground truth due to the use of RGB or RGB-D inputs for annotations.

III. Methodology

SUPER considers the problem of estimating upper body human poses when a subject faces mmWave radar sensors at a known distance. The assumption for known distance is valid in confined environments such as in an office cubicle or inside a

[1] Demonstration videos can be found at https://super-2023-web.github.io/SUPER/.



(a) Camera view.

(b) OptiTrack motion capture view.

(c) SUPER output

Fig. 1: The estimated skeleton model from SUPER vs. ground truth when a subject raises her/his hand up while seating. The blue circle markers stand for the estimated skeleton model, and the red plus markers are the corresponding ground truth.

TABLE I: Comparison of existing works on mmWave-based HPE

Method	Radar Sensor	Ground Truth Sensor	Data Representation	Body Motions
mm-Pose [13]	2 TI AWR1642	Microsoft Kinect	two 2D intensity image (XY-plane and XZ plane)	Walking Left-Arm Swing, Right-Arm Swing, Both-Arms-Swing
MARS [28]	1 TI IWR1443	Microsoft Kinect	5D Point Cloud (x, y, z, velocity, intensity)	10 rehabilitation movements^1
mmMesh [14]	1 TI AWR1843	VICON system	6D Point Cloud (x,y,z,range, velocity, energy)	8 daily activities^2
m^4esh [30]	1 TI AWR1843	VICON system	6D Point Cloud (x, y, z, range, velocity, energy)	7 daily activities^{3†}
HuPr [31]	2 TI IWR1843	RGB camera	VRDAEMap(velocity-specific range-doppler-azimuth-elevation map)	freely performed by multi-person static actions, standing and waving hand(s), walking with waving hand(s)
mmBody [32]	Arbe Robotics Phoenix	MoCap system	6D dense Point Cloud (x, y, z, velocity, amplitude, energy)	100 motions
Ours	2 TI IWR6843	OptiTrack system	intensity point cloud, Doppler point cloud	upper limb movements, head rotation, driving simulation

[1] Right/left/both limb extension, right/left side lunge right/left front lunge, right/left upper body extension, squad.

[2] Torso rotations, clockwise walking, counter-clockwise walking, arm swing), walking back and forth; walking back, and forth with arm swing, walking in the place, lunges.

[3] 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.

[†] Freely performed by multi-person in one recording.

vehicle. Alternatively, existing approaches for mmWave-based target localization can be adopted to determine a bounding box around the subject [33]. In this section, we first provide an overview and the design rationale of the SUPER pipeline and then present details of its individual components.

A. Overview and Design Rationale

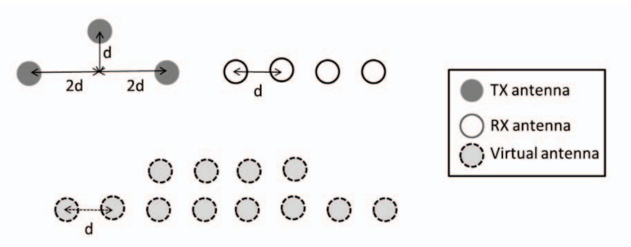


Fig. 2: A 2-Dimensional MIMO antenna array for IWR6843ISK radar. The separation d equals half wavelength.

Low-end COTS mmWave radars typically have a small number of Tx and Rx antennas, which restrict their spatial resolution. Take TI IWR6843ISK radar as an example. It features 3 Tx and 4 Rx antennas forming a 12 virtual antenna array as illustrated in Figure 2. Placed horizontally, this configuration results in angle resolutions of 15 degrees and 55 degrees, respectively, in the horizontal and vertical directions. To estimate fine-grained SUB-HPE, a high azimuth angle resolution is necessary for extremities when the arms are extended while a high elevation angle resolution is helpful in distinguishing subtle head and trunk poses. To mitigate the limitations of low-end mmWave radars, we employ two closely located radar sensors: one oriented horizontally and the other vertically. Despite the lack of coordination, the reflected wave from one radar's transmission is unlikely mistaken as that from the other radar since the resulting range bins are outside the region of interest (ROI). Note that although dual-radar systems have been also employed in mm-Pose [13] and HuPR [31], the data is used to produce 2D heatmaps (images) in perpendicular planes rather than being fused together in 3D point clouds.

Several existing mmWave-based HPE methods model human body as a point cloud, which is obtained from range-Doppler maps over multiple chirps of radar signals. Doppler information has sufficient coverage on the entire body only if there are significant motions in different body parts. In seated positions, however, movements in the trunk and low limbs are confined leading to sparse points in space. In contrast, the intensity of reflected signals from the bulk of the body is high regardless of motions as long as the subject is sufficiently close to the radars. Thus, a range-angle map, augmented with intensity information from a multi-antenna system, better captures the occupation of human body in space. Motivated by this observation and with the unique characteristics of seated SUB-HPE in mind, we extract two point clouds with reflected intensity and Doppler information. The ablation study in Section V further substantiates the empirical evidence supporting the complementary nature of the two input sources.

The overall system diagram of SUPER shown in Figure 3, consists of two main processing blocks, i.e., point cloud generation and a backbone network. The reflected RF signals from two radar sensors are preprocessed using match filtering and range-FFT. Dense point clouds are then generated by sampling the ROIs in 3D space centred around each radar. A dual-radar fusion algorithm coherently combines data from two radars and samples the results to produce fine-grained point cloud data representation for intensity and for Doppler. Both point clouds are fed into the backbone network. The network comprises building blocks from PointNet [34], PointNet++ [35], and LSTM to extract global and local features to predict the SMPL pose parameters in each frame. The pipeline can be easily extended to predict body shape parameters and will be investigated as part of future work.

B. Point cloud generation with dual-radar fusion

In this section, we introduce a novel pipeline to generate quality point clouds from data collected by two closely located radars. Data from each radar goes through separate branches to handle intensity and Doppler information. The overall processing consists of two stages: the first stage transforms raw radar data to a dense point cloud, which acts as an intermediate representation. In the second stage, data from the two radar sensors are fused together and then sampled to produce a fine-grained point cloud.

1) Dense point cloud generation: Raw I-Q samples from each radar in intermediate frequency (IF) follow the standard pre-processing steps. These include mapping the raw radar data into a range map through range-FFT and DC compensation to eliminate static background clutters. As previously mentioned, SUPER operates under the assumption that the approximate distance between the subject and the radars is known. This knowledge enables the designation of an ROI that encompasses the subject. For example, when seated around 1 meter away from the radars, the range bins that span the subject's body are approximately from 0.4 meters to 1.8 meters. These parameters can be easily adjusted given the setup of different scenarios.

Intensity point clouds: To generate intensity point clouds, we further consider 180-degree field of view (FOV) in both horizontal and vertical directions and choose a non-uniform sampling scheme as shown in Table II. Specifically, for the radar placed horizontally (radar H), θ and ϕ correspond to the azimuth and elevation angles; for radar V, the reverse is true. Clearly, as indicated in Table II, angles are densely sampled in the axis where more spaced virtual antennas are available and near the center, while in the perpendicular direction, fewer angle bins are sampled. Consequently, amongst the 30 range bins between 0.4 meters and 1.8 meters from the subject, there are in total 6930 (= $21 \times 11 \times 30$) sample points in the ROI.

TABLE II: Non-uniform Angle Sampling (unit in degree)

θ	-70	-60	-50	-40	-30	-25	-20	-15	-10	-5	0
	5	10	15	20	25	30	40	50	60	70	
ϕ	-70	-50	-30	-20	-10	0	10	20	30	50	70

Next, we apply Minimum Variance Distortionless Response (MVDR) to generate an intensity spectrum for each point location in the ROI. We first estimate the correlation matrix for each range index i, using all N chirps within one frame,

$$R_i = \frac{\sum_{n=1}^{N} \mathbf{y}\mathbf{y}^H}{N},$$
$$R_i = R_i + \alpha \frac{trace(R_i)}{K} I_K,$$

where $\mathbf{y}$ is a column vector of the received signal at each antenna, N is the number of chirps in one frame, K is the number of received antennas, and α is a control parameter to prevent singularity.

Next, we calculate the steering vector $\mathbf{a_s}$ from the virtual antennas array as

$$\mathbf{a_s}(n) = \begin{cases} exp(j\pi(\mu_a(n-1))), & 1 \leq n \leq 8, \\ exp(j\pi(\mu_a(n-6-1)+\mu_b)), & 9 \leq x \leq 12, \end{cases}$$

where

$$\mu_a = sin(\theta\pi/180)cos(\phi\pi/180),$$
$$\mu_b = sin(\phi\pi/180).$$

Finally, we calculate the intensity spectrum for each sample point as

$$IS(\theta, \phi, i) = \frac{1}{\mathbf{a_s}^H R_i^{-1} \mathbf{a_s}},$$

where $\mathbf{a_s}$ is the steering vector, and i is the range index. This process creates a 4D point cloud with intensity values in polar coordinates, which can then be transformed into a dense point cloud in a Cartesian coordinate system centered on a radar.

Doppler point clouds: To generate Doppler point clouds, we follow a similar procedure to that in [14]. Specifically, Doppler-FFT on the chirps in a frame is applied to derive 2D range-Doppler maps (30×128) of each received antenna. For every point in the 2D range-Doppler map, its velocity and power are calculated through an additional angle-FFT across multiple received antennas. The procedure is applied to data from the two radars independently, resulting two 5D point clouds of 3840 (= 128×30) points for each radar.

It is worth noting that the term "dense" is adopted to differentiate this representation from the eventual fused point

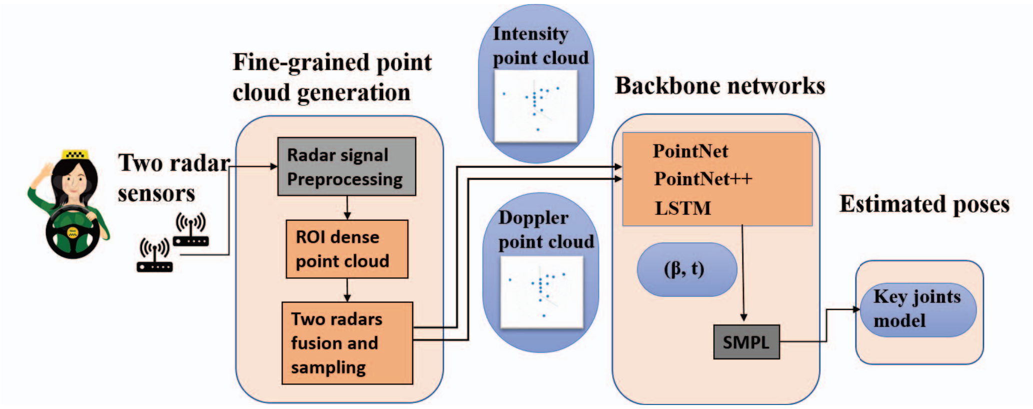


Fig. 3: The system diagram of SUPER. New processing blocks introduced in this paper are highlighted in orange, and intermediate data flows are highlighted in blue.

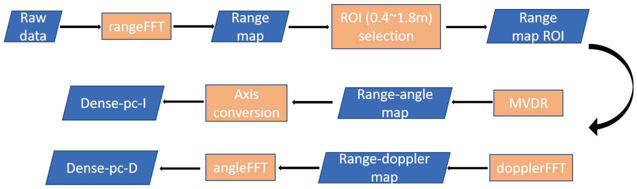


Fig. 4: Generation of dense point clouds from raw radar data. One intensity point cloud and one Doppler point cloud are produced for each radar separately.

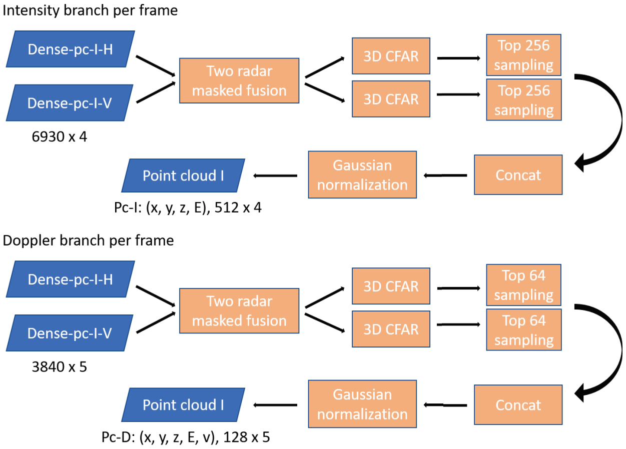


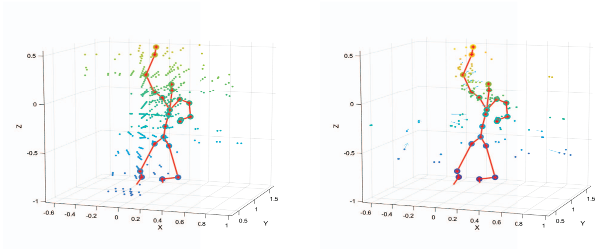
Fig. 5: Generation of fine-grained point clouds by fusing and sampling dense point clouds from the two radars.

clouds. While the point clouds in this initial stage remain relatively sparse when compared to those generated by Lidar sensors, they are denser than the point clouds typically found in existing literature on mmWave-based HPE. This increased density is achieved through spatial oversampling in the intensity point clouds. Further information regarding the process is illustrated in Figure 4.

2) Dual-radar fusion for fine-grained point clouds: To this end, we have generated four point clouds, i.e., one 4D intensity point cloud and one 5D Doppler point cloud from each radar. The two radar sensors are positioned in close proximity, approximately 15cm apart. Thus, the dense point clouds generated by each radar sensor roughly share the same ROI but are complementary spatially. Radar H captures detailed information in the horizontal direction, which can be used to enhance the quality of the point cloud derived from radar V, and vice versa. Therefore, the purpose of dual-radar fusion is two-folded. First, it refines the point clouds from one radar using the point clouds from the other radar. Second, it trims the over-sampled point clouds and retains only salient points. At the end of the procedure, a single intensity point cloud and a single Doppler point cloud are obtained for further processing. An overview of this process is given in Figure 5.

Masked refinement: To refine the point clouds from both radars, we first transform their representations from polar coordinate frames to a unified Cartesian coordinate frame. Consider the 4D intensity point clouds from radar H as an example. A similar procedure is applied to the intensity point cloud from radar V and the 5D Doppler point clouds from both radars. Let the point cloud from radar H be the target and that from radar V serves as a reference. For each point in the target point cloud, the K nearest points in the reference point cloud are identified. The mean power value of these points is computed through averaging. The value of the point in the target point cloud is replaced by the product of itself and the mean value. This multiplication has the effect of masking or suppressing points with high values in only one point cloud and amplifying those with high values in both. Furthermore, the operation can preserve local power variations, as the masks within the same local area are nearly identical.

Point cloud trimming: Due to spatial over-sampling, the dense point clouds produced thus far contain redundant information. To retain only informative points, we extend the principles of the 2D Constant False Alarm Rate (CFAR) algorithm [36] and implement a 3D CFAR algorithm, by



(a) An intensity point cloud.

(b) A Doppler point cloud.

Fig. 6: An example fine-grained point clouds. Ground truth skeleton is shown in red. The magnitude and direction of Doppler velocity are shown in arrows

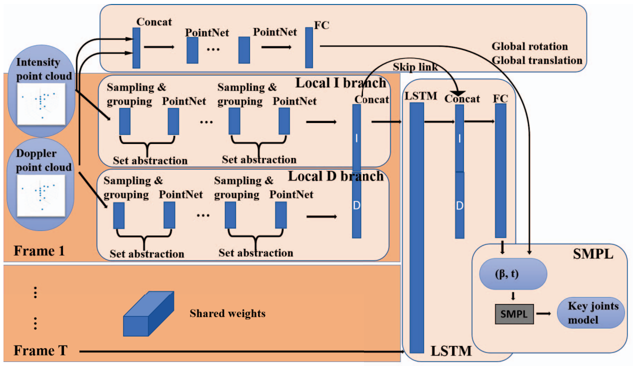


Fig. 7: The architecture of the deep neural network backbone.

adaptively calculating thresholds to detect local peaks as key points. Finally, we output the top 256 key points from the intensity point clouds and the top 64 key points for the Doppler point clouds.

Following the extraction of key points, we merge the point clouds from both radars and apply a Gaussian normalization filter to the values. The final fine-grained point cloud consists of 512 key points, featuring $[x, y, z, intensity]$ for intensity, and 128 key points with attributes $[x, y, z, power, velocity]$ for Doppler. An example fine-grained point clouds generated from the process is shown in Figure 6. In this example, the subject raises their right hand to the top. It is evident from Figure 6a, the intensity points are present not only around the raised arm but also at other areas of the upper body. In contrast, as shown in Figure 6b, the Doppler points mainly appear around the raising arm with non-negligible velocity.

C. The deep neural network backbone

A deep neural network (Figure 7) is designed to take multiple frames of fine-grained point clouds as inputs to predict joint positions in a human skeleton model. The network incorporates both global and local contexts to estimate the intricate translation and rotation dynamics. To capture the global context, we include a dedicated branch that stacks three basic PointNet blocks [34]. To extract local information, three hierarchy set abstraction layers in PointNet++ are stacked to process both the intensity and Doppler point clouds [35].

Furthermore, to exploit the temporal dependencies between frames, two layers of unidirectional Long Short-Term Memory (LSTM) cells are used [37], spanning $T = 20$ steps or frames (equivalent to one second). To enhance information flow, a skip/residual link is introduced that connects features prior to the LSTM layers and post-LSTM. Finally, after several fully connected (FC) layers, the model outputs rotations of each joint within the human skeleton model. To improve the accuracy of rotation estimation, following [38], we represent joint rotations using 6D parameters of the rotation matrices rather than 3D axis angles.

The model subsequently leverages SMPL to generate the final joint positions. A gender-neutral model is used by fixing the default shape parameters. For seated upper body poses, we freeze the rotation parameters of joints in the lower body and only estimate the positions of the upper body joints (14 joints) [7].

The loss function is defined as the mean square error (MSE) of the joint coordinates:

$$Loss = \frac{1}{F} \sum_{f=1}^{F} ||P_{f,J}^{(f)} - P_{gt,J}^{(f)}||_2, \tag{1}$$

where f is the frame index, F is the total number of frames in the batch, J denotes the joint set, $P_{f,J}^{(f)}$ is the estimated positions of key joints, and $P_{gt,J}^{(f)}$ is the corresponding ground truth positions. Note that the loss is a function of the pose parameters (β) and global translation (t). From the experiments, we find that instead of directly regressing the joint positions, passing the joint rotation parameters through SMPL to estimate the resulting joint position errors results in higher accuracy and faster convergence. This can be interpreted as a non-linear transformation of the MSE loss function using the SMPL model.

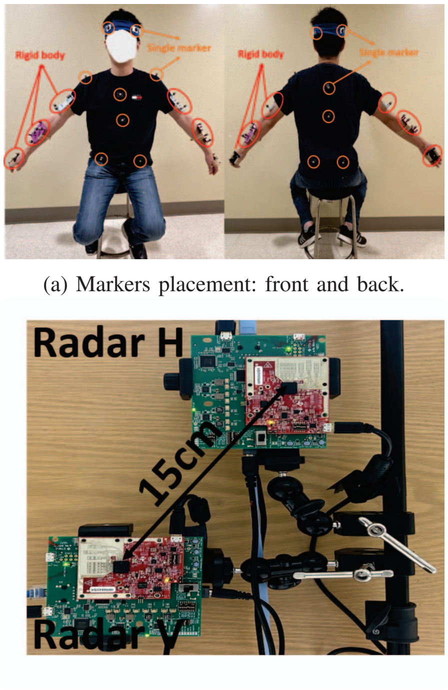
The total number of learning parameters in the network is 2.9 million or 2.65G FLOPs. Incoming point clouds are processed in a sliding window manner with a window size of 20 frames.

IV. Implementation and Datasets

In this section, we present the implementation of a prototype SUPER system using COTS mmWave radars and the experiment results from multi-subject testbed evaluations under various conditions, which are purposely chosen to closely mimic real-life situations.

A. Implementation

Two IWR6843ISK boards [39] together with DCA1000EVM boards [40] are used in the experiments. The radar boards operate at 60 ∼ 64 GHz (with 4-GHz bandwidth) and transmit FMCW signals. The radar front-ends include 3 transmit antennas (Tx), 4 receive antennas (Rx), with 120° azimuth field of view (FoV) and elevation FoV. The 3 transmitting antennas emit FMCW chirps in a time-division



(a) Markers placement: front and back.

(b) Co-located radar sensors.

Fig. 8: Experiment setup: markers and radars.

TABLE III: Radar Hardware Settings.

parameters	description	values
N_{tx}	number of transmit antennas	3
N_{rx}	number of receive antennas	4
$N_{virtual}$	number of virtual antennas	12
P_f	frame duration	50 (ms)
f_s	start frequency	60 (GHz)
f_e	end frequency	64 (GHz)
t_{rs}	start ramp time	0 (μs)
t_{re}	end ramp time	58 (μs)
t_{idle}	chirp idle time	7 (μs)
N_{adc}	number of samples per chirp	225
N_{chirp}	number of chirps per frame	128

manner, which results in a 12 virtual antennas array. Each FMCW chirp is composed of 225 sampling points, and the frequency of RF will increase from 60 GHz to 64 GHz. 128 chirps constitute one frame at a frame rate of 20Hz. The acquired raw IF signal is sent to a host PC via Ethernet, where mmWave Studio [41] is used to initiate, configure, and control the radar boards. The detailed radar sensor settings is summarized in Table III.

The preprocessing steps and point cloud generation are implemented in MATLAB R2021a, which takes raw IF signals as input, and outputs the fine-grained 3D point cloud data. The neural network backbone is implemented in PyTorch.

B. Data collection procedure

To evaluate SUPER's performance, we recruited 10 participants (3 females and 7 males), aged between 21 and 46, and with BMI in the range of $18.1 \sim 31.6$. Participants wore their daily attire such as T-shirts, blouses, and sweaters of different fabric materials. This research protocol has been approved by the research ethical board (REB) from our institution.

Both radar and mocap data are collected in a $6.5m \times 6m$ lab. The lab (Figure 9) has standard office furniture and many electronic equipment and wireless transceivers (WiFi, LTE, Bluetooth, etc.). Both radar sensors on a tripod as in Figure 8b with 1.5 meters high and 1 meter away from the subjects and oriented at a 20-degree horizontal angle. We define a local coordinate system with respect to radar H. During the experiments, only one subject is present in the predefined position.

Ground truth of subject poses are collected from OptiTrack, a motion capture system [42] with 12 cameras. Both radar sensors and the OptiTrack system are synchronized after data collections at frame level using "synchronization" motions at the beginning of each trial. The output of the OptiTrack system are coordinates of markers and rigid bodies on the body of participants as shown in Figure 8a. We utilize MotionBuilder [43] to build a customized human actor for each participant and generate accurate joints coordinates through motion tracking functionalities built in the software. Videos have been recorded during data collection for reviewing purposes but are not further processed.

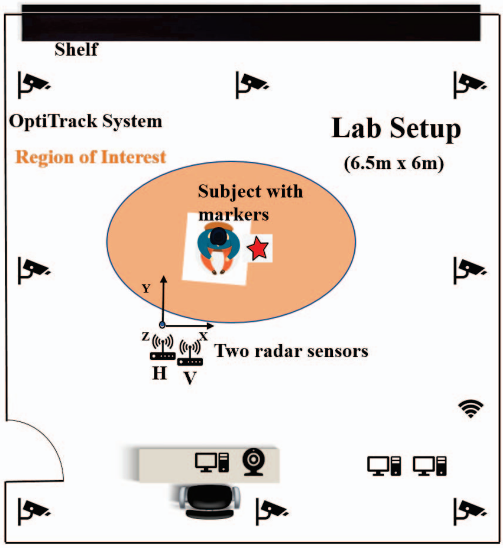


Fig. 9: The Lab environment for data collection.

During the data collection process in the controlled laboratory environment, subjects engaged in three distinct motion sequences that are designed to mimic movements while seated in confined environments. These include: hand-reaching, driving, and head rotation. A Microsoft Xbox Gaming steering wheel is used to mimic a driving platform and is placed in front of the subjects.

- *Hand-reaching trials:* Participants were instructed to use their right hand to interact with hypothetical objects in their surroundings while keeping their left hand stationary. These trials included interacting with objects positioned directly above one's head (top), in the top front (up-front), in front but to the side (right-front), to the side (right), and below (bottom).
- *Driving trials:* These trials aimed to replicate common driving activities. Subjects were instructed to perform routine driving (with both hands on the wheel), conduct traffic checks (by leaning forward and inspecting both left and right directions), engage in a conversation with a passenger (rotating the head towards the passenger), execute reverse maneuvers (turning the head to see one's back over the shoulder), and operate the control panel (reaching the right-front area and virtually press buttons with one's right hand).
- *Head rotation trials:* These trials capture deliberate head movements while keeping one's torso mostly stationary. Subjects were instructed to look left and right, up and down, and upper/lower left/right, etc.

C. The dataset

In total, we conducted 30 trials from 10 participants, with each lasting around 10 minutes. The total number of radar frames collected is around 360,000 from each radar sensor. The total size of all raw radar data in the dataset is around 900GB. The ground truth data for each frame contains joint angles and positions of 14 upper body key joints[2] and the global translations. The total size of the ground truth data is around 900MB. The dataset is organized by subject ID (de-identified), trial name, and data types (radar data vs ground truth data).

V. Performance Evaluation

In this section, we present the performance of SUPER and ablation studies.

A. Evaluation metrics and baseline method

We chose three metrics in literature to quantify the accuracy of the estimated upper body joints. The first one is the Mean Per Joint Position Error (MPJPE), which measures the *absolute* average distance (mm) between the predicted joints of a human skeleton and the ground truth joints in a given dataset. The MPJPE is defined as:

$$E_{MPJPE}(f, J) = \frac{1}{K_J} \sum_{k=1}^{K} ||P_{f,J}^{(f)}(k) - P_{gt,J}^{(f)}(k)||_2, \qquad (2)$$

where f denotes a frame, J denotes the joints model/set, K is the number of joints in the model/set, $P_{f,J}^{(f)}(k)$ is the estimated position of joint k, and $P_{gt,J}^{(f)}(k)$ is the corresponding ground truth position. Finally, the MPJPEs are averaged over all frames.

TABLE IV: Accuracy of joint estimations (in mm)

action	method	MPJPE↓	PA_MPJPE↓	PCK@15mm↑
driving	mmMesh	156.85±25.18	29.60±6.2	13.76±7.38
	Ours	**112.46±12.70**	**16.32±2.45**	**27.38±11.15**
handreaching	mmMesh	148.33±25.18	26.97±4.39	15.74±8.65
	Ours	**114.87±25.07**	**15.19±2.56**	**37.17±8.28**
head rot.	mmMesh	174.42±40.61	30.43±8.82	10.00±6.85
	Ours	**108.85±15.46**	**16.16±2.59**	**28.46 ±9.34**

The second metric is the Procrustes alignment MPJPE (PA-MPJPE) that calculates the average 3D joint distance (mm) after performing Procrustes alignment [44] on the estimated and ground-truth joint sets. PA-MPJPE measures how well the pose estimation model captures the structural information of the pose, rather than just its location or scale. It eliminates system biases and allows for fair comparisons across different scales of the same pose.

The third metric is the percentage of correct keypoints under a distance threshold e.g. $15mm$ ($PCK@15mm$). This metric is defined as:

$$PCK@15mm = \frac{1}{K} \sum_{k=1}^{K} \delta_k, \qquad (3)$$

where K is the total number of keypoints (joints), δ_k is a binary value indicating whether the distance between the ground truth keypoint and the predicted keypoint is within a certain threshold.

Baseline method: We adopt mmMesh [14] as the baseline method for comparison. The choice is primarily driven by the fact that the model architecture was made publicly available by the authors. Although m^4esh is more recent, it targets multi-subject scenarios, which are outside the scope of this paper. The mmMesh model parameters were retrained using the hyperparameters suggested in [14] and data from radar H only for consistency.

B. Main results

We calculate the average MPJPE, PA_MPJPE, and PCK@15mm of the 14 upper body joints in *leave-one-subject-out experiments*. The results presented in TABLE IV reveal that our approach remarkably surpasses the baseline model by average margins(take the average of the three actions) of 30%, 45%, and 184% on MPJPE, PA_MPJPE, PCK@15mm respectively.

Furthermore, we evaluate the model's effectiveness on upper limb joints pivotal to hand-object interactions. The the MPJPE and PA_MPJPE of left and right wrist and elbow joints are summarized in Table V.

From the results, we can see that the average accuracy of wrist joints is lower than that of elbow and other upper body joints. This can be explained by the low RCS of hands, making them difficult to be captured by radars. However, SUPER considerably outperforms mmMesh in the estimation of both

[2] The local body joints are set to fixed sitting poses

TABLE V: Accuracy of upper limb key joint positions (in mm)

action	method	MPJPE↓		PA_MPJPE↓	
		wrist	elbow	wrist	elbow
driving	mmMesh	341.96±64.83	199.94±37.81	103.35±25.63	66.50±18.48
	Ours	**119.46±25.71**	**114.30±14.16**	**38.95±6.98**	**28.45±6.20**
handreaching	mmMesh	312.22±66.61	187.68±30.50	96.44±21,47	64.44±16.39
	Ours	**140.46±30.77**	**124.86±27.79**	**42.57±11.65**	**32.30±9.78**
head	mmMesh	377.14±102.65	226.84±56.17	104.40±29.49	68.90±20.96
	Ours	**131.08±26.88**	**127.10±31.05**	**38.70±9.41**	**27.96±6.84**

upper arm joints. Thus, we conclude that it is important to design a specific pipeline for SUB-HPE, and the inclusion of intensity features and the use of radar V are instrumental in improving the accuracy.

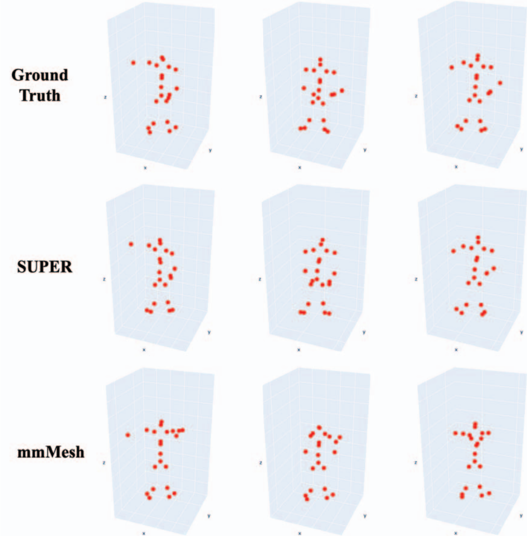


Fig. 10: Constructed 3D poses in skeleton representation from SUPER, mmMesh and Ground Truth

Examples of constructed 3D poses in skeleton representations using SUPER, mmMesh and Ground Truth can be found in Fig 10.

C. Ablation study

TABLE VI: Results from Ablation Study

action	information	MPJPE↓	PA_MPJPE↓	PCK@15mm↑
driving	Doppler only	237.33±26.72	44.78±2.29	0.37±1.46
	intensity only	196.97±32.89	20.09±6.94	25.36±15.98
	Doppler+intensity	101.51±19.06	12.27±3.41	47.40±17.18
handreaching	Doppler only	266.28±36.64	48.49±3.74	0.08±0.91
	intensity only	193.05±43.37	19.26±7.94	26.75±19.83
	Doppler+intensity	110.73±30.91	12.51±5.86	47.10±20.53
head	Doppler only	221.46±31.73	46.04±1.21	0.02±0.63
	intensity only	190.71±44.41	19.05±7.47	27.68±18.90
	Doppler+intensity	99.67±28.14	12.28±5.87	45.25±18.79

* The standard deviation in this table is calculated across the estimated position errors per joint and per frame.

We further conduct ablation experiments to evaluate the effectiveness of Doppler and intensity point clouds in the training data. To do so, we only input the Doppler point cloud or the intensity point cloud and remove the respective branch in the backbone (Figure 7). Table VI reports the results from one test subject performing different actions. Clearly, neither intensity or Doppler point clouds alone is sufficient. Combining both sets of features leads to the highest accuracy. Somewhat interesting, between the two, intensity point clouds appear to be more informative.

VI. Demonstrative Application

In this section, we demonstrate the utility of SUPER through a downstream task that identifies hand-object interaction through SUB-HPE. Note that what is being presented acts as a proof-of-concept. Likely, more sophisticated methods can be implemented for the task on top of SUB-HPE.

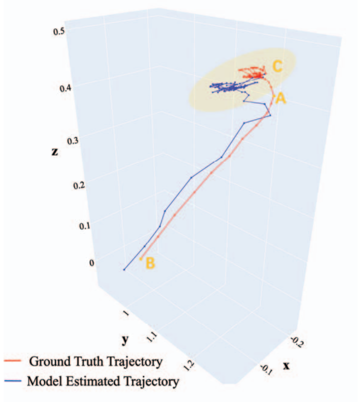


Fig. 11: Visualization of the ground truth and model estimated wrist trajectories of a 4s sequence from a handreaching action. The units are in meters.

In this task, the aim is to determine which objects in the 3D space one is interacting with by hands. Consider a motion sequence where a hand starts from some resting position, moves toward an object at known location, and then interacts with the object for a period of time. We transform the problem of object identification to a localization problem, namely, to determine whether one's hand (a wrist joint specifically) falls into the predefined bounding boxes around target locations for a sufficient amount of time.

To test this idea, we first calculate the amount of displacement of a wrist joint during 1s windows in the ground truth trajectory. The intervals that the total displacement is less than a predefined threshold (100mm in the implementation) indicate either the initial rest position or the rendezvous point between the hand and a target object. We compute the centroid of the wrist joint positions in such intervals and test against the ground truth target locations. As an example, consider the ground truth and estimated trajectories as shown in Figure 11. In this example, one's hand travels from B to A and then reaches a target location C. Although the estimated trajectory does not exactly coincide with the ground truth one, it can be observed as the hand approaches and stays around the target location, the estimated locations are close to C.

We conduct experiments on all subjects using the hand-reaching trials. The results show that in 88.80% of the rest position or the rendezvous point intervals, the centroid of the estimated writ trajectory falls into a bounding box centered on the target location with a side length of 0.2m.

VII. Discussion and Conclusion

In this work, we proposed SUPER, a pipeline for SUB-HPE. To address the challenges of nuanced upper body movements when seated, we obtained both intensity and Doppler point clouds by fusing data coherently from two radars with orthogonal orientations. Compared to a baseline method that only utilizes Doppler point clouds from a single radar, SUPER has superior performance in terms of all metrics for HPE.

The current SUPER framework assumes the presence of a single subject and the knowledge of the ROI. It can be easily extended to multiple subjects and unknown ROIs when combined with a target detection component. The current model can also be trained with additional mesh errors in SMPL and a term reflecting temporal consistency and smoothness of human movements [45]. Doing so is expected to further improve the accuracy and realism of the inferred poses.

Future research directions for mmWave-based SUB-HPE also include developing models that are robust to different deployment environments and the investigation of more downstream tasks.

References

[1] Y. Liu, J. Yang, X. Gu, Y. Guo, and G.-Z. Yang, "Ego+x: An egocentric vision system for global 3d human pose estimation and social interaction characterization," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 5271–5277.

[2] J. Wang, K. Qiu, H. Peng, J. Fu, and J. Zhu, "Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance," ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3343031.3350910

[3] T. Anvari and K. Park, "3d human body pose estimation in virtual reality: A survey," in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 2022, pp. 624–628.

[4] Y. Zhou, H. Huang, S. Yuan, H. Zou, L. Xie, and J. Yang, "Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation," *IEEE Internet of Things Journal*, vol. 10, no. 16, pp. 14 128–14 136, 2023.

[5] S. Y. Cheng and M. Trivedi, "Turn-intent analysis using body pose for intelligent driver assistance," *IEEE Pervasive Computing*, vol. 5, no. 4, pp. 28–37, 2006.

[6] C. E. Matthews, S. A. Carlson, P. F. Saint-Maurice, S. Patel, E. Salerno, E. Loftfield, R. P. Troiano, J. E. Fulton, J. N. Sampson, C. Tribby *et al.*, "Sedentary behavior in united states adults: Fall 2019," *Medicine and science in sports and exercise*, vol. 53, no. 12, p. 2512, 2021.

[7] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, oct 2015. [Online]. Available: https://doi.org/10.1145/2816795.2818013

[8] A. A. Osman, T. Bolkart, and M. J. Black, "Star: Sparse trained articulated human body regressor," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 598–613.

[9] Z. Cao, G. H. Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[10] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, 2022.

[11] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–15, 2018.

[12] V. Mollyn, R. Arakawa, M. Goel, C. Harrison, and K. Ahuja, "Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–12.

[13] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[14] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su, "mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 269–282.

[15] J. Geng, D. Huang, and F. De la Torre, "Densepose from wifi," *arXiv preprint arXiv:2301.00250*, 2022.

[16] P. Hügler, M. Geiger, and C. Waldschmidt, "Rcs measurements of a human hand for radar-based gesture recognition at e-band," in *2016 German Microwave Conference (GeMiC)*. IEEE, 2016, pp. 259–262.

[17] T. Gu, Z. Fang, Z. Yang, P. Hu, and P. Mohapatra, "Mmsense: Multi-person detection and identification via mmwave sensing," in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 2019, pp. 45–50.

[18] C. Wu, F. Zhang, B. Wang, and K. R. Liu, "mmtrack: Passive multi-person localization using commodity millimeter wave radio," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 2400–2409.

[19] P. Zhao, C. X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "mid: Tracking and identifying people with millimeter wave radar," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2019, pp. 33–40.

[20] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–19, 2016.

[21] H. Liu, A. Zhou, Z. Dong, Y. Sun, J. Zhang, L. Liu, H. Ma, J. Liu, and N. Yang, "M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3397–3415, 2021.

[22] S. Palipana, D. Salami, L. A. Leiva, and S. Sigg, "Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 5, no. 1, pp. 1–27, 2021.

[23] A. Khamis, B. Kusy, C. T. Chou, M.-L. McLaws, and W. Hu, "Rfwash: a weakly supervised tracking of hand hygiene technique," in *Proceedings of the 18th conference on embedded networked sensor systems*, 2020, pp. 572–584.

[24] Z. Yang, P. H. Pathak, Y. Zeng, X. Liran, and P. Mohapatra, "Monitoring vital signs using millimeter wave," in *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*, 2016, pp. 211–220.

[25] P. Zhao, C. X. Lu, B. Wang, C. Chen, L. Xie, M. Wang, N. Trigoni, and A. Markham, "Heart rate sensing with a robot mounted mmwave radar," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2812–2818.

[26] F. Wang, X. Zeng, C. Wu, B. Wang, and K. R. Liu, "mmhrv: Contactless heart rate variability monitoring using millimeter-wave radio," *IEEE Internet of Things Journal*, vol. 8, no. 22, pp. 16 623–16 636, 2021.

[27] B. Zhang, B. Jiang, R. Zheng, X. Zhang, J. Li, and Q. Xu, "Pi-vimo: Physiology-inspired robust vital sign monitoring using mmwave radars," *ACM Transactions on Internet of Things*, vol. 4, no. 2, pp. 1–27, 2023.

[28] S. An and U. Y. Ogras, "Mars: mmwave-based assistive rehabilitation system for smart healthcare," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 20, no. 5s, pp. 1–22, 2021.

[29] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[30] H. Xue, Q. Cao, Y. Ju, H. Hu, H. Wang, A. Zhang, and L. Su, "M4esh: mmwave-based 3d human mesh construction for multiple subjects," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 391–406.

[31] S.-P. Lee, N. P. Kini, W.-H. Peng, C.-W. Ma, and J.-N. Hwang, "Hupr: A benchmark for human pose estimation using millimeter wave radar," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5715–5724.

[32] A. Chen, X. Wang, S. Zhu, Y. Li, J. Chen, and Q. Ye, "mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3501–3510.

[33] W. Chen, H. Yang, X. Bi, R. Zheng, F. Zhang, P. Bao, Z. Chang, X. Ma, and D. Zhang, "Environment-aware multi-person tracking in indoor environments with mmwave radars," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 3, sep 2023.

[34] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85.

[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5105–5114.

[36] M. A. Richards, *Fundamentals of Radar Signal Processing, 2nd Edition*. McGraw Hill, 2005.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, nov 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[38] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5738–5746.

[39] T. Instruments, "Iwr6843isk," 2020. [Online]. Available: https://www.ti.com/product/IWR6843

[40] ——, "Dca1000evm," 2020. [Online]. Available: https://www.ti.com/tool/DCA1000EVM

[41] ——, "mmwave studio," 2020. [Online]. Available: http://www.ti.com/tool/MMWAVE-STUDIO

[42] OptiTrack, "Optitrack: Motion capture systems," 2020. [Online]. Available: https://www.optitrack.com/

[43] AUTODESK, "Motionbuilder," 2022. [Online]. Available: https://www.autodesk.com/

[44] J. Gower, "Generalized procrustes analysis." *Psychometrika*, vol. 40, p. 33–51, 1975.

[45] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2023.